ELIMINATING RATER BIASES IN PERFORMANCE APPRAISAL

Purnell H. Benson
Rutgers University

Sources of Error in Evaluating Performance with Rating Scales

The use of rating scales to measure the performance of individuals raises questions of sources of errors and methods for detecting and controlling the errors. In using a rating scale where a numbered step or numerical point is selected to judge performance, various types of error appear.

(1) The constant bias of the individual rater who habitually overrates or underrates all ratees is a bias which shifts the origin or zero point of the scale used. If the correct mean for a set of individuals rated by rater $\underline{k}$ is $\bar{X}_k'$, and the mean of the ratings reported by rater $\underline{k}$ is $\bar{X}_k$, the two values are related by $\bar{X}_k = \bar{X}_k' + Z_k$, where Z_k is the bias or shift in the zero point of the scale resulting from $\underline{k}$'s constant error in judgment.

(2) The habitual contraction or expansion in the dispersion of ratings by rater $\underline{k}$ is a distortion introduced by those who are either reluctant to give extreme ratings or who go to extremes in choosing ratings. If the standard deviation of ratings by $\underline{k}$ is SD_k, and the correct standard deviation for those ratings is SD_k', the two are related by a stretch correction factor F_k such that $SD_k' = F_k SD_k$.

(3) Also involved is the interpersonal error which rater $\underline{k}$ makes with regard to ratee $\underline{i}$. This is an error unique to ratee $\underline{i}$ and rater $\underline{k}$, designated as P_{ik}.

(4) Remaining is a residual random error which depends upon the precision of judgment of which rater $\underline{k}$ is capable, E_k. In practice, P_{ik} may not be separable from E_k, and the two may be considered together as a residual error R_k characteristic of rater $\underline{k}$.

Combining the three types of error into a single measurement equation, the correct rating of individual $\underline{i}$ is related to the rating given of individual $\underline{i}$ by rater $\underline{k}$ by

$$X_i' = (X_{ik} - \bar{X}_k) \cdot F_k + \bar{X}_k - Z_k - R_k, \qquad (1)$$

where $\bar{X}_k$ is the mean of the ratings made by rater $\underline{k}$ of $\underline{k}$'s ratees, F_k is the stretch correction for rater $\underline{k}$, Z_k is the average bias in the ratings by $\underline{k}$, and R_k is the residual error for rater $\underline{k}$.

Computation to Obtain Scale Values for Ratee Performance Which Eliminate Zero-Point Biases of Raters

With an array of ratings X_{ik} of ratees $\underline{i}$ by raters $\underline{k}$, we seek to use the information contained in this matrix to learn the correct ratings X_i' of the performance of the individual ratees. This involves removing the zero-point and stretch errors. We first consider eliminating the zero-point biases of the raters. Reduction of the residual error to a minimum random error of judgment will be considered later.

The matrix of ratees rated by raters yields intervals between the performance ratings of all pairs of ratees from the sets of ratees with common raters. Each rating interval is correct in the sense that the constant zero-point bias has been subtracted in defining the rating interval. If these intervals are correct except for a random residual error, they can be averaged by the arithmetic of paired comparisons to obtain more accurate estimates of the rating intervals between all pairs of ratees from the entire group of ratees.

We proceed as follows. For each pair of ratees $\underline{i}$ and $\underline{j}$ for whom rating intervals are given by one or more raters $\underline{k}$, we average the $\underline{q}$ intervals to obtain for the pair of ratees $\underline{i}$ and $\underline{j}$:

$$Y_{ij} = \sum_{k=1,q} (X_{ik} - X_{jk})/q. \qquad (2)$$

The average interval Y_{ij} is posted in the cell for the $\underline{i}$th column and the $\underline{j}$th row of a paired matrix, and again with the sign reversed in the cell for the $\underline{j}$th column and the $\underline{i}$th row. If the data yield pair differences for all possible pairs of ratees in the matrix, the numerical average of the pair differences down the $\underline{i}$th column gives the average interval between ratee $\underline{i}$ and all of the ratees included by the matrix.

If, as is often the situation, the matrix is incompletely paired, the regression procedure reported by F. Mosteller (1951) is used to find scale values whose differences provide the best fit in the least squares sense to the pair differences which are included in the incomplete matrix. The input for the regression calculation consists of 1 and -1 in the $\underline{i}$th and $\underline{j}$th columns of the row with Y_{ij} as the entry for the dependent variable. Entries elsewhere are 0's, except for the last row which contains 1's to establish the origin for the system of scale values. The entry for the dependent variable in this added row is 0.

The sums of cross-products and squares are calculated about an origin of 0, rather than the mean. This reflects the circumstance that only the entries from one side of the diagonal of the paired matrix need be used in the calculation. The matrix of squares and cross-products whose solution yields the scale values for performance contains entries as follows, using W_{ij} as the weight for the number of raters who define the interval between ratees $\underline{i}$ and $\underline{j}$.

The diagonal cells of the $\underline{i}$th row and column contain $\sum_{j=1,n} W_{ij} + 1$. The off-diagonal cells for the $\underline{i}$th row and $\underline{j}$th column have $1 - W_{ij}$. The $\underline{i}$th row of the n+1 column for dependent variable is $\sum_{j=1,n} W_{ij} Y_{ij}$. The sum of squares for the dependent variable is $\sum_{i=1,n-1} \sum_{j=1,n} W_{ij} Y_{ij}^2$.

The scale values S_i found from solving the equations in this matrix are performance ratings about a mean of zero. While they define rating intervals between ratees, they are not performance ratings in an absolute sense. The norm for the group of ratees must be known, so that the scale values can be transformed to this norm.

The norm may be defined according to some external behavior criterion, or it may be fixed by expert judgment, or it may be taken as the simple average of all of the ratings of ratees by raters. The proper performance norm is the one which is meaningful to those using the rating scale for which a norm is needed.

Calculation to Eliminate the Stretch Bias of Raters

We now consider removal of the stretch bias evident in the contraction or expansion of ratings by each rater. The stretch differences of individual raters can be made uniform by imposing the same rating dispersion upon all raters. Of course it is necessary to consider that each rater may rate a somewhat different set of ratees. First, the standard deviation of performance ratings calculated for all ratees is fixed. Then, the spread of ratings by each rater is altered to agree with the spread of ratings calculated for that rater's ratees. This provides an adjusted set of ratings by each rater for a next iteration of computation. Iteration continues until no further adjustment in the spread of any rater's ratings takes place.

The rating X_{ik}' for ratee $\underline{i}$ and rater $\underline{k}$ corrected for the stretch in the rater's scale is related to the unadjusted rating X_{ik} by

$$X_{ik}' = F_k(X_{ik} - \overline{X}_k) + \overline{X}_k, \qquad (3)$$

where $\overline{X}_k$ is the mean of the original ratings made by rater $\underline{k}$ of rater $\underline{k}$'s ratees, and F_k is a stretch correction factor defined by

$$F_k = SD_k'/SD_k, \qquad (4)$$

with SD_k the standard deviation of rater $\underline{k}$'s ratings, and SD_k' is the standard deviation of $\underline{k}$'s ratees obtained from the ratings calculated for these ratees.

Like the mean imposed as the correct norm upon the system of performance ratings, the standard deviation designated can be defined by an external behavior criterion, or by expert judgment, or simply taken as the standard deviation of all ratings by all raters.

If the standard deviation of the performance scale values X_i' calculated from the inputted ratings is SD', and these scale values are about a mean of zero, then the scale values X_i'' transformed to a designated mean $\overline{X}_o$ and standard deviation SD_o are given by

$$X_i'' = \overline{X}_o + (SD_o/SD')X_i' \qquad (5)$$

Control of Interpersonal and Random Errors in Rating

No simple computational procedure is available to eliminate the interpersonal error peculiar to a particular rater and ratee. This type of error arises from favoritism and misjudgment of the unique achievements of the ratee. As for the random error remaining, this is an error unrelated to systematic analysis.

Both of these types of residual error depend upon the ability and motivation of the rater to control them. Instruction of raters in the criteria for making ratings is important in reducing residual errors, as well as zero-point and stretch errors. Improvement in precision of judgment requires measurement of rating accuracy to grant recognition to those who are efficient raters. If those who play favorites or who fail to take the rating effort seriously are detected by having their rating efficiency measured, this affords means for improving rating efficiency or avoiding those whose rating activity is of poor quality.

Several components of rating efficiency can be isolated and measured by comparing ratings made by raters with the ratings calculated from input by competent raters. We will call these calculated ratings "adjusted group ratings" or AGR. Various comparisons of original ratings with the calculated ratings yield scores.

(1) The zero-point score of rater k, referred to as score ${}_1T_k$, can be defined as follows for m items of performance rated:

$$ {}_1T_k = \left[{}_1\bar{Z}' - \left[\sum_{h=1,m} \frac{(\bar{X}_{h.k} - \bar{X}_{h.k}')^2}{m} \right]^{\frac{1}{2}} \right] \cdot \left[\frac{{}_1SD_o}{{}_1SD} \right] + {}_1\bar{X}_o, \qquad (6) $$

where ${}_1\bar{X}_o$ is the mean imposed upon performance scores, ${}_1SD_o$ is the standard deviation imposed, ${}_1\bar{Z}'$ is the mean zero-point bias of raters. (Rater bias is a standard deviation of item biases of each rater). ${}_1SD$ is the standard deviation of the zero-point biases of raters (calculated as standard deviations of item biases), $\bar{X}_{h.k}$ is the mean rating by rater k of ratees for performance item h, and $\bar{X}_{h.k}'$ is the mean AGR calculated for rater k on item h.

(2) The score ${}_2T_k$ for stretch bias of rater k as a standard deviation of item differences from the AGR spread for k's ratees is

$$ {}_2T_k = \left[{}_2\bar{Z}' - \left[\sum_{h=1,m} \frac{(SD_{h.k} - SD_{h.k}')^2}{m} \right]^{\frac{1}{2}} \right] \cdot \left[\frac{{}_2SD_o}{{}_1SD} \right] + {}_2\bar{T}_o, \qquad (7) $$

with the same identification of variables as before, except for the prescript 2 reference to stretch bias.

(3) The score ${}_3T_k$ for residual error of rater k after adjusting k's ratings for zero-point and stretch biases depends first upon the calculation for each item of the residual standard error. This is calculated with $n - 2$ degrees of freedom (or $n - 1$ if k has only 2 ratees, not permitting a valid adjustment for stretch). Then the root mean square of the residual standard errors is obtained with m degrees of freedom for m items.

$$ {}_3T_k = \left[{}_3\bar{R}' - \left[\sum_{h=1,m} \frac{(R_{h.k})^2}{m} \right]^{\frac{1}{2}} \right] \left[\frac{{}_3SD_o}{SD_1} \right] + {}_3\bar{T}_o, \qquad (8) $$

where $R_{h.k}$ is the residual error of rater k rating item h.

(4) If provision is made for raters to make self-ratings, the discrepancy between the self-rating and AGR, rather than the residual error after adjusting the self-rating for zero-point bias and stretch bias.

$$ {}_4T_k = \left[{}_4\bar{Z}' - \left[\sum_{h=1,m} (X_{h.kk} - X_{h.k}')^2 \right]^{\frac{1}{2}} \right] \cdot \left[\frac{{}_4SD_o}{{}_4SD} \right] + {}_4\bar{T}_o, \qquad (9) $$

where $X_{h.kk}$ is the self-rating by rater k on item h, $X_{h.k}'$ is the AGR for ratee k, ${}_4\bar{Z}'$ is the mean self-rating error on all items (expressed as a standard deviation), and ${}_4SD_o$ and ${}_4T_o$ are the standard deviation and mean imposed for self-rating scores.

In practice, the conversion of error quantities into standard scores is more simply accomplished if the four separate error quantities are averaged into a single score for rating efficiency. Then the mean and standard deviation are imposed upon the overall rater score. Each separate rater score has subtracted from it the group mean for that type of score and then is divided by the group standard deviation for that type of score. This converts all four error scores to the same standard deviation. Then the average of the four scores for each rater is calculated, and a group standard deviation for the overall rater scores is calculated. With the ratio of this to the imposed standard deviation used as a multiplier of the divergence of the overall score from the group mean, in the same manner as the separate formulas already given, the overall scores are converted to those with the required mean and standard deviation. The formula for the combined score adjusted to the imposed standard deviation and mean for the group is applied to

$$ T_k = {}_1T_k + {}_2T_k + {}_3T_k + {}_4T_k, \qquad (10) $$

if equal weights are assigned to each of the four error components with unit standard deviations, and the final formula for adjustment is

$$ T_k' = (T_k - \bar{T}) \left[\frac{SD_o}{SD} \right] + \bar{T}_o, \qquad (11) $$

with T_k' the final rater score, SD the standard deviation of T_k for the group, and $\bar{T}$ the mean of the T_k before adjustment, and $\bar{T}_o$ the mean rater score imposed for the group.

Comparison of ratings by each rater with those made by the leader for that rater's ratees permit four more error scores to be defined. The overall score can be made a weighted combination of the two sets of four scores, if all are available.

Since these measures of rating efficiency depend upon the difficulty of the task of rating, some adjustment is needed when poor performers are

rated who cannot be rated with the same absolute precision as good performers who are near the top of the rating scale.

In the PEERRATE system, diminution of the measure of rater error is accomplished by one of the following two formulas.

$$R_{i.k}' = R_{i.k}\left[\frac{a_3}{a_0 - a_1 X_i'}\right], \qquad (12)$$

$$R_{i.k}' = R_{i.k}\left[a_4 . 10^{a_2(a_1 X_i')}\right]. \qquad (13)$$

$R_{i.k}'$ is the error after adjusting the residual error $R_{i.k}$ in the rating of ratee $\underline{i}$ by rater $\underline{k}$, and a_0, a_1, a_2, a_3 and a_4 are parameters found suitable for the error adjustment. Such parameters can be selected to maximize the correlation between the score for rater efficiency and some criterion, such as the rating received for performance on items.

The rater's score for rating efficiency and the same rater's performance score as a ratee can be used to calculate a suitable weight in the calculation of the adjusted group rating AGR. Commencing with equal weights for raters, these can be progressively improved through iteration, using fresh weights at each stage of iteration obtained from the rater scores from the previous stage of iteration. In the PEERRATE system, the performance score and rater score are combined by parameters for linear, square and cross-product terms.

Operation of the PEERRATE Computer Program

The PEERRATE rating system described here has been implemented by a computer program prepared by the author of the system. An early version of the program was reported by Benson (1976).

The computer program permits a variety of computations to be made to meet various rating situations, such as use or non-use of leader ratings, use of team or department ratings, and combination of ratings into rating scores by either addition or multiplication of ratings together. The program also calculates a matrix of intercorrelations between the performance and rating scores, item by item or overall scores. These intercorrelations help guide the operator towards the selection of proper parameters for the calculations made. All of the results and intermediate steps of calculation can be outputted on cards, tape or disk, at the option of the user of the program, to facilitate further research.

The PEERRATE program, consisting of a deck of approximately 3,000 cards, is available on application to: Dean Horace J. De Podwin, Graduate School of Business Administration, Rutgers University, Newark, N. J. 07102. The program is free of cost to educational and non-profit users except for cost of transcribing the program on cards or tape.

Tables 1, 2, & 3 contain inputted ratings and calculated scores for item performance and rating efficiency.

References

Benson, P.H. A computerized system of student grading of student assignments. Third Annual New Jersey Conference on the Use of Computers in Higher Education, Rutgers University, New Brunswick, N.J., March 22, 1976, pp. 21-30.

Mosteller, Frederick. Remarks on the method of paired comparisons: the least squares solution assuming equal standard deviations and equal correlations. Psychometrika, 16 (March 1951, pp. 3-9.

Table 1

Rater-Ratee Matrix of Ratings for Items

Rater	Item	1	2	3	4	5	6	7	8	9	10	11	12
1	1	***	80	80	87	***	73	87	67	67	93	73	80
	2	***	67	100	67	***	53	80	47	100	73	53	80
2	1	***	87	93	93	***	80	80	73	73	87	93	93
	2	***	87	87	93	***	80	80	73	80	80	100	93
3	1	***	87	100	80	***	67	93	80	67	87	100	87
	2	***	80	93	93	***	67	93	67	80	87	80	87
4	1	***	73	73	80	***	27	67	53	80	73	67	67
	2	***	67	47	87	***	33	87	67	87	73	67	80
5	1	***	93	***	***	***	***	***	***	67	87	***	***
	2	***	73	***	***	***	***	***	***	93	87	***	***
6	1	***	100	93	100	***	100	100	87	87	93	93	93
	2	***	87	87	93	***	100	100	93	87	100	93	93
7	1	***	80	80	87	***	60	87	73	73	87	93	73
	2	***	67	93	73	***	67	87	73	80	73	87	73
8	1	***	***	87	87	***	87	80	87	***	***	93	87
	2	***	***	87	87	***	93	87	87	***	***	93	87
9	1	***	80	73	87	***	93	87	93	87	73	93	93
	2	***	87	93	93	***	93	93	100	93	87	100	93
10	1	***	87	87	87	***	87	93	73	80	100	87	87
	2	***	87	87	87	***	80	93	80	93	93	93	93
11	1	***	***	53	67	***	***	80	***	***	***	73	93
	2	***	***	93	93	***	***	53	***	***	***	100	80
12	1	***	100	100	93	***	93	93	93	87	93	100	93
	2	***	87	100	93	***	93	93	93	87	93	100	93

Table 2

Ratee Performance Scores

	Rating by Group				Rating by Leader			
Id. No.	Rater Quality	Overall Rating	Item 1 Rating	Item 2 Rating	Rater Quality	Overall Rating	Item 1 Rating	Item 2 Rating
1	***	***	***	***	***	***	***	***
2	91	82	89	76	100	74	80	67
3	82	90	88	92	100	90	80	100
4	92	91	92	90	100	77	87	67
5	***	***	***	***	***	***	***	***
6	70	70	71	68	100	63	73	53
7	85	91	91	92	100	84	87	80
8	83	73	73	73	100	57	67	47
9	78	82	73	92	100	84	67	100
10	87	88	91	86	100	83	93	73
11	88	92	94	91	100	63	73	53
12	96	89	89	90	100	80	80	80

Table 3

Rater Performance Scores

		Comparison With Group Calculations				Comparison With Leader Rating			
Id. No.	Overall Rater Score	Average Deviation	Difference In Range	Residual Error	Self Deviation	Average Deviation	Difference In Range	Residual Error	Self Deviation
1	100	100	100	***	***	***	***	***	***
2	90	99	96	85	90	87	88	83	90
3	93	96	93	92	88	91	92	99	90
4	86	70	74	84	88	90	100	89	90
5	98	96	90	100	***	100	100	99	***
6	75	79	87	73	70	70	76	75	70
7	95	84	97	90	98	100	91	96	100
8	76	93	80	70	78	76	70	70	71
9	80	86	89	70	86	75	74	71	90
10	89	93	90	89	89	81	77	99	90
11	81	70	70	100	75	93	95	76	70
12	82	79	84	79	100	70	72	80	92