DOCUMENT RESUME

ED 229 420 TM 830 319

AUTHOR	Littlefield, John H.; And Others
TITLE	Adjusting Observational Ratings to Improve Inter-Rater Consistency.
PUB DATE	15 Apr 83
NOTE	12p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE	Speeches/Confererce Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Bias; *Error of Measurement; Higher Education; *Interrater Reliability; *Medical Evaluation; Medical School Faculty; Medical Schools; Medical Students; *Observation; *Scoring Formulas; *Student Evaluation

ABSTRACT

Observational ratings of student clinical performance are influenced by factors other than the quality of the performance. Individual raters may be more stringent or lenient than their colleagues. In this medical school setting, multiple raters evaluated each student. To reduce the influence of "error" due to differences among raters, each rater was assigned a handicap score which was calculated in three steps: (1) identify the cohort of students observed by the rater, (2) calculate the mean of all faculty ratings for that cohort (grand mean) and the mean given those students by the rater, and (3) subtract the individual rater mean from the grand mean. Analysis of the "original" and "adjusted" ratings for two academic years indicated no differences in overall mean and standard deviation. Generalizability analysis indicated an improvement equivalent to increasing the number of raters per student by 50 percent (i.e., the variance component due to error was reduced by about 33 percent). (Author)

ADJUSTING OBSERVATIONAL RATINGS TO IMPROVE INTER-RATER CONSISTENCY

John H. Littlefield, Ph.D.
Nancy E. Anthracite, M. D.
Robert Herbert, M.S.
Jean McKendree, B.S.

UNIVERSITY OF TEXAS HEALTH SCIENCE CENTER AT SAN ANTONIO

Paper Presented at the Annual Meeting of the
American Educational Research Association
Montreal, Canada
April 15, 1983

PRINTED IN U.S.A.

ADJUSTING OBSERVATIONAL RATINGS TO IMPROVE INTER-RATER CONSISTENCY

John H. Littlefield, Ph.D., Nancy E. Anthracite, M.D., Robert Herbert, M.S.,
and Jean McKendree, B.S.-University of Texas Health Science Center at San Antonio

Introduction

Observational ratings are a widely used method for assessing student clinical performance in health science education. Common measurement errors associated with rating forms include errors of leniency and central tendency, halo effect, logical error, proximity error and contrast error. (DeMers, 1978) Wherry (1952) discusses rating errors by using an equation to picture the complexity of the rater's task. The recorded rating score can be represented by the following equation:

$$RS = (A + e_b) + (E + e_e) + (B + e_p) + e_r$$

In this equation, RS is the recorded score, A is the ability of the student, E is environmental influence, B is the bias of the rater, and e_x represents errors due to atypical behavior of the student, unexpected changes in the environment, aberrant perceptions by the rater and random fluctuations respectively. In classical test theory terminology, A is equivalent to a true score. E represents the influence of environmental factors such as the format of the rating form, training and motivation levels of the raters, and the performance situations in which students are observed. B represents bias due to the idiosyncracies of an individual rater. This study reports a procedure for adjusting recorded scores to reduce the influence of rater bias. If the numerical size of $(B + e_\rho)$ is reduced in the equation above, RS will be a more accurate estimate of the student's ability level (A).

Nunnally (1978) points out that raters differ in leniency, the tendency to say good or bad things about people in general. In an educational context, students would describe raters with leniency errors as "tough" or "easy" graders. Littlefield, et. al., (1981a) demonstrated that differences in rater leniency of medical faculty were constant over a five year period despite annual comparative feedback to the faculty. Cason and Cason (1981) propose a construct called Rater Reference Point to account for individual differences in rater leniency. The Cason model uses latent trait theory to estimate each rater's reference point (i.e., leniency error). This report proposes a similar adjustment to ratings by individual faculty raters; however, instead of latent trait theory, individual faculty are assigned a "handicap" score based upon the mean of all of the various faculty ratings given to the students observed by the individual rater.

Method

The subjects in this study are 203 medical faculty and residents who rated at least five junior medical students during a 3 1/2 week Internal Medicine Clerkship in academic years 1981 and 1982. The requirement to have rated at least five students was arbitrarily imposed to insure that each rater had performed sufficient ratings to establish a "stable mean." The design of the rating form and the role of attending faculty have been described previously. (Littlefield, et. al., 1981b). Performance was rated on each of five items on a 0-to-14 point numerical scale.

A total of 355 medical students were each rated by 5 to 9 raters during academic years 1981 and 1982. A "handicap" score was calculated for each "subject" rater in three steps: 1) identify the cohort of students rated by the individual faculty member during the academic year (range = 5 to 49); 2) calculate the mean of all faculty ratings for that cohort (grand mean) and the mean rating given those students by the individual faculty member and 3) subtract the individual rater mean from the grand mean ($h=x_g-x_r$). If a rater received a positive "handicap" score, his/her mean score was lower than the grand mean for all raters who observed that cohort of students. All individual faculty ratings were "adjusted" by adding the handicap score to the original ratings. The result was two sets of ratings for each student, original and adjusted. The data sets were edited using two criteria: (1) eliminate student records which do not have at least four ratings by "subject" raters and (2) randomly delete ratings from student records with more than four. The final result was two 4 X 162 matrices (adjusted and unadjusted student ratings) for 1981 and two 4 X 144 matrices for 1982. Generalizability analyses (Brennan and Kane, 1977) were performed on the original and the adjusted ratings. This analysis uses an analogy to communications systems to assess the precision of the scores. The variance component due to differences between students (the signal) is compared to the variance component due to differences among raters of the same student (noise). Variance components are statistical estimates of the hypothesized components of an observed score (Cronbach, et. al., 1972). The numerical size of the variance component due to differences between students is directly related to the standard deviation of the mean rating given to each student. The numerical size of the variance component due to differences between raters of the same student is directly related to how closely the four raters agree. The BMDP-8V program (Dixon and Brown, 1979) was used to compute variance components.

Results

Table 1 reports the overall mean ratings, standard deviations and range for the original and adjusted ratings in academic years 1981 and 1982. It appears that the adjustments did not substantially change the overall leniency of the ratings or the "spread" among individual student scores. Table 2 presents a frequency distribution of the number of faculty with various levels of "handicap" scores. A Kolmogrov Smirnoff test indicates that the handicap scores in 1981 and 1982 approximate a normal distribution. Like many human traits, a few individuals apparently have rather extreme positive or negative leniency error, but most raters are near zero. The overall means of the handicap scores are 0.04 in 1981 and 1982 with standard deviations of .851 and .843 respectively. Table 3 is an analysis of variance summary table for the original and adjusted ratings. Notice in the adjusted ratings that the sums of squares due to differences between raters of the same student decrease substantially from the original ratings. This would be expected since the handicap score adjusts each individual faculty's ratings toward the grand mean for the cohort of students rated.

Table 4 presents the intraclass correlation coefficients for the original and adjusted ratings. These coefficients summarize the ability of the ratings to separate the "signal," in this case the differences (variance) among students, from the "noise." The coefficients can vary from 0.0 to 1.0. The 1981 original ratings coefficient is in the same range as those reported by Littlefield, et. al. (1981b) when adjusted to reflect four raters. The 1982 original ratings coefficient is higher due to a larger signal (variance component due to differences between students). The adjusted ratings coefficients are larger than the original ratings coefficient due to an increase in the strength of the "signal" and a decrease in the "noise." The decrease in "noise" reflects the reduced mean square due to differences between raters of the same student. The increased "signal" strength is also related to reduced noise since it is calculated by the expected mean square (EMS) equation: EMS=$4\sigma^2_{st} + \sigma^2_{r(s)}$. In this equation EMS is set equal to the mean square due to differences between students. $\sigma^2_{r(s)}$, the variance component due to differences between raters of the same student is set equal to its analogous mean square. With estimates of EMS and $\sigma^2_{r}(s)$, the equation can then be solved to find $\sigma^2{}_{st}$. Table 5 demonstrates the algebraic manipulation. Table 6 demonstrates the change of individual student scores for the 1981 academic year.

Discussion and Conclusions

This study has implications in two areas: making decisions about students based upon observational ratings and improving the precision of ratings. A large clinical department utilizes many faculty raters and they are likely to differ in leniency error. With relatively random assignment of students to raters, some students will be assigned entirely to stringent raters and their mean rating score in this system (0-14 scale) will be one to two points lower than their performance level justified. The seriousness of this problem depends on the types of decisions to be made. In this particular system, it might result in the change of a letter grade, but not in outright failure because failure decisions are reviewed individually by the clerkship director. Table 6 shows that only 55% of the students in 1981 would remain in the same decile as their unadjusted mean rating score. The changes in decile for students in 1982 were not calculated; however, they would be less pronounced because the variance component due to students (signal) is much larger indicating that the scores are more spread out.

Adjusting rating scores is an inexpensive method of improving the precision of rating systems. Landy and Farr (1980) in a review of the research on performance ratings note that training raters will reduce rating errors if the training is sufficiently extensive. In this rating system, over 200 raters observe the students therefore the logistics of training raters are formidable if not prohibitively difficult. By contrast, the use of "adjusted scores" represents an improvement in precision of the scores with no additional costs. The degree of improvement will depend on the relative strength of the "signal" and "noise" variance components. The improvement in the 1982 intraclass correlation coefficient was less striking than in 1981. From an organizational development perspective, it would be critically important to involve the raters in the decision to adopt adjusted scores. The validity of the ratings depends upon the conscientious efforts of the raters and the validity of the whole process would suffer immensely if they are trying to "beat the system."

This study has demonstrated a method for estimating the effects of rater bias on recorded scores as outlined by the equation:

$$RS = (A+e_b) + (E+e_e) + (B+e_p) + e_r$$

The handicap scores are a composite estimate of $B+e_p$ Each rater submitted only one rating per student, therefore, random errors of perception cannot be separated from the effects of overall bias (B). The findings of the study must be qualified by noting that the requirement that subject raters have completed at least five ratings resulted in deleting about 50% of the raters from each academic year. Five ratings established a "stable mean" for each rater from which his/her handicap score could be calculated. It seems likely that raters who complete less than five ratings annually are more susceptible to leniency error than their colleagues who rate larger number of students. Landy and Farr (1980) emphasize the need to learn more about the way raters observe, encode, store, retrieve and record performance information. With that research, perhaps answers will come to questions such as the accuracy of ratings by "occasional" raters.

Bibliography

American Psychological Association, Standards for Educational and Psychological Tests, Washington, D.C., American Psychological Association, 1974 p. 49.

Brennan, R. and Kane, M., Signal/noise ratios for domain-referenced tests, Psychometrika, V.42: 609-625, 1977; Errata, 43: 289, 1978.

Cason, G. and Cason, C., Some promising early results from a rudimentary latent-trait theory of performance rating paper presented at the annual meeting of AERA, Los Angeles, 1981 ERIC Document ED-201-669.

Cronbach, L., Gleser, G., Nanda, H. and Rajaratnam, N., The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles, New York: J. Wiley and Sons, 1972 p. 25.

DeMers, J., Observational Assessment of Performance, In Morgan, M. and Irby, D. Evaluating Clinical Competence in the Health Professions, St. Louis: C. V. Mosby, 1978.

Dixon, W. and Brown, M. BMDP-79, Berkeley: University of California Press, 1979.

Landy, F. J. and Farr, J. L., Performance rating, Psychological Bulletin, V. 87, (1), p. 72-107, 1980.

Littlefield, J., Anthracite, N., Kromer, M. and Harrington, J., A longitudinal study of observational ratings by clinical faculty. Proceedings of the Twentieth Annual Conference on Research in Medical Education, Association of American Medical Colleges, 1981a.

Littlefield, J., Harrington, J., Anthracite, N., and Garman, R., A description and four-year analysis of a clinical clerkship evaluation system, Journal of Medical Education, V. 56, p. 334-340, 1981b.

Nunnally, J., Psychometric Theory, New York: McGraw-Hill, 1978, p. 563.

Wherry, R. J., The control of bias in rating: A theory of rating (Personnel Research Board Rep. 922) Washington, D.C.: Department of the Army Personnel Research Section, Feb., 1952.

TABLE 1

Descriptive Statistics for Original and Adjusted Ratings

1981		1982	
Original	Adjusted	Original	Adjusted
$x = 9.17$	$x = 9.26$	$x = 9.29$	$x = 9.33$
$\sigma = 1.18$	$\sigma = 1.16$	$\sigma = 1.91$	$\sigma = 1.68$
Range=6.2-12.4	Range=5.97-12.11	Range=3.0-14.0	Range=4.13-14.87

TABLE 2
Frequency Distribution of Handicap Scores

1981 Score Range	No. of Faculty Raters	1982 Score Range	No. of Faculty Raters
-1.99 to -1.50	3	-2.41 to -1.50	5
-1.49 to -1.00	14	-1.49 to -1.00	7
- .99 to - .50	13	- .99 to - .50	13
- .49 to 0.00	13	- .49 to 0.00	22
.01 to .49	25	.01 to .49	25
.50 to .99	21	.50 to .99	18
1.00 to 1.49	10	1.00 to 1.49	10
1.50 to 2.01	3	1.50 to 2.01	3
	100		103

TABLE 3
ANOVA Summary Tables & Variance Components for Original and Adjusted Ratings

		Source	Sum of Squares	D.F.	Mean Square	Variance Component
1981	Adjusted Ratings	Students	862.76	161	5.36	.92
		Difference Bet.Raters	821.50	486	1.69	1.69
	Original Ratings	Students	898.76	161	5.58	.68
		Difference Bet.Raters	1392.01	486	2.86	2.86
1982	Adjusted Ratings	Students	915.63	143	6.40	1.19
		Differences Bet.Raters	708.25	432	1.64	1.64
	Original Ratings	Students	1007.2	143	7.04	1.13
		Difference Bet.Raters	1096.2	432	2.54	2.54

TABLE 4
Intraclass Correlation Coefficient for Original and Adjusted Ratings

	1981	1982
Conceptual Model	$\rho = \frac{\text{signal}}{\text{signal + noise}}$	$\rho = \frac{\text{signal}}{\text{signal + noise}}$
Adjusted Ratings	$\rho = \frac{.92}{.92 + 1.69/4} = .69$	$\rho = \frac{1.19}{1.19 + 1.64/4} = .74$
Original Ratings	$\rho = \frac{.68}{.68 + 2.86/4} = .49$	$\rho = \frac{1.13}{1.13 + 2.54/4} = .64$

TABLE 5
Calculating the Variance Component Due to Student Differences

$$EMS = 4\,\sigma^2_{st} + \sigma^2_{r(st)}$$

$$MS_{st} = 4\,\sigma^2_{st} + MS_{r(st)}$$

$$\sigma^2_{st} = \frac{MS_{st} - MS_{r(st)}}{4}$$

TABLE 6
Impact on Decisions about Students in 1981
Class Quartile Changes
(N = 162)

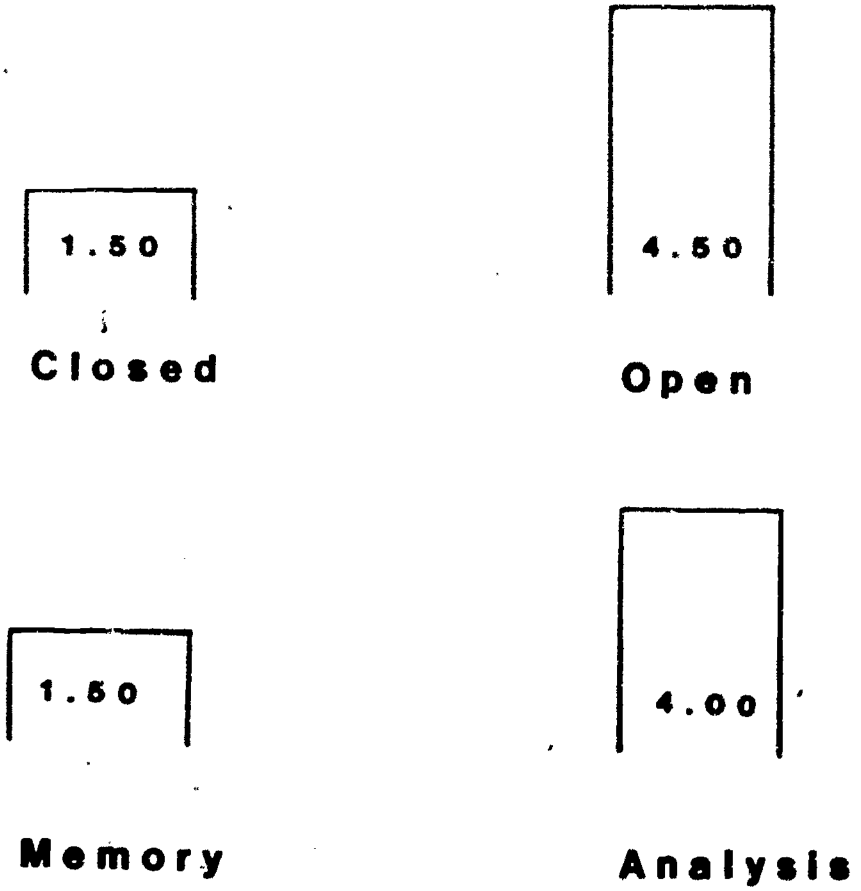
Down	No Change	Up
9%	83%	8%

Class Decile Changes

(N = 162)

Down 2	Down 1	No Change	Up 1	Up 2
2%	22%	55%	19%	2%

EFFECT OF OPEN AND CLOSED QUESTIONS ON PARTICIPATION



Andrews, Teaching Development Newsletter, 1980

* Types of Lead off Questions

TYPE:	RESPONSE:
Quiz Show	1.50
Fishing	2.00
Shotgun	2.50
Metaphysical (General invitation)	2.50
Structured Open	5.00

CLOSED QUESTIONS

characteristics:

- predictable answers
- tests memory of student
- often yes/no, or one word answer
- will not stimulate discussion

Abstract

ADJUSTING OBSERVATIONAL RATINGS TO IMPROVE INTER-RATER CONSISTENCY

Observational ratings of student clinical performance are influenced by factors other than the quality of the performance. Individual raters may be more stringent or lenient than their colleagues. In this medical school setting, multiple raters evaluated each student. To reduce the influence of "error" due to differences among raters, each rater was assigned a handicap score which was caluculated in three steps: (1) identify the cohort of students observed by the rater, (2) calculate the mean of all faculty ratings for that cohort (grand mean) and the mean given those students by the rater, and (3) subtract the individual rater mean from the grand mean. Analysis of the "original" and "adjusted" ratings for two academic years indicated no differences in overall mean and standard deviation. Generalizability analysis indicated an improvement equivalent to increasing the number of raters per student by 50 percent (i.e., the variance component due to error was reduced by about 33%).