DOCUMENT RESUME

ED 260 104 TM 850 426

AUTHOR	Littlefield, John H.; And Others
TITLE	Metacognition of Performance Raters.
PUE DATE	Apr 85
NOTE	7p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Analysis of Variance; *Clinical Experience; Confidence Testing; *Evaluators; Graduate Medical Education; Higher Education; *Interrater Reliability; Medical School Faculty; *Medical Students; Metacognition; *Rating Scales; *Student Evaluation; Teacher Student Relationship; Test Reliability; Test Validity

AESTRACT

Sixteen Family Practice faculty members completed ratings on 59 senior medical students after a 6-week primary care clerkship. Each student was rated by seven to ten faculty members and the chief residents who worked with them, resulting in a total of 353 ratings. The rating scale covered: (1) attainment of learning objectives; (2) progress during the clerkship; (3) overall performance, (4) frequency of contact between student and rater (number of patients discussed); and (5) confidence in the rating, to indicate raters' metacognition. A two-factor analysis of variance was performed on the results to explore the relationships among rater accuracy, level of contact, and rater confidence in the score assigned. It was concluded that confidence in the validity of a rating was not related to the accuracy of that rating. Level of rater-student contact was, however, related to accuracy, with the most accurate ratings based upon discussion of seven to eleven patients. Low levels of contact were associated with overly stringent ratings, and high levels of contact were associated with lenient ratings. Individual raters differed in the leniency of scores, the tendency to make extreme judgments, and confidence in each rating. (GDC)

Metacognition of Performance Raters

John H. Littlefield, Ph.D.

Richard E. Ellis, M.D.

Robert J. Herbert, M.S.

Peter A. Cohen, Ph. D.

University of Texas Health Science Center
at San Antonio

Paper Presented at the Annual Meeting of the
American Educational Research Association
Chicago, Illinois
April, 1985

Metacognition of Performance Raters

John H. Littlefield, Ph.D.. Richard E. Ellis, M.D.,
Robert J. Herbert, M.S. and Peter A. Cohen, Ph.D.

University of Texas Health Science Center at San Antonio

INTRODUCTION

Recent theoretical work regarding performance ratings identifies two separate bodies of research, one concerned with the rating instrument and the other with the cognitive processes of the rater (Feldman, 1981). In a review of performance rating research, Landy and Farr (1980) recommend a moratorium on rating form research until we learn more about how raters observe, encode. store, retrieve and record performance information. This study assesses the accuracy of individual judgements, then examines its relationship with rater confidence in those judgements and the level of rater-ratee contact. The primary goal is to better understand situational factors which are related to the accuracy of performance ratings.

The existence of long term consistent differences in leniency error of clinical performance raters has been recently reported (Marienfeld & Reid, 1984; Littlefield et. al., 1984). These individual differences can be reduced by statistically adjusting the scores (Littlefield, et. al., 1984, Cason & Cason, 1984). This solution is less than satisfactory because it may reduce student confidence in the validity of feedback from faculty raters and may also reduce faculty motivation to provide accurate ratings. A preferable solution would be to modify rater idiosyncracies to produce more standardized ratings.

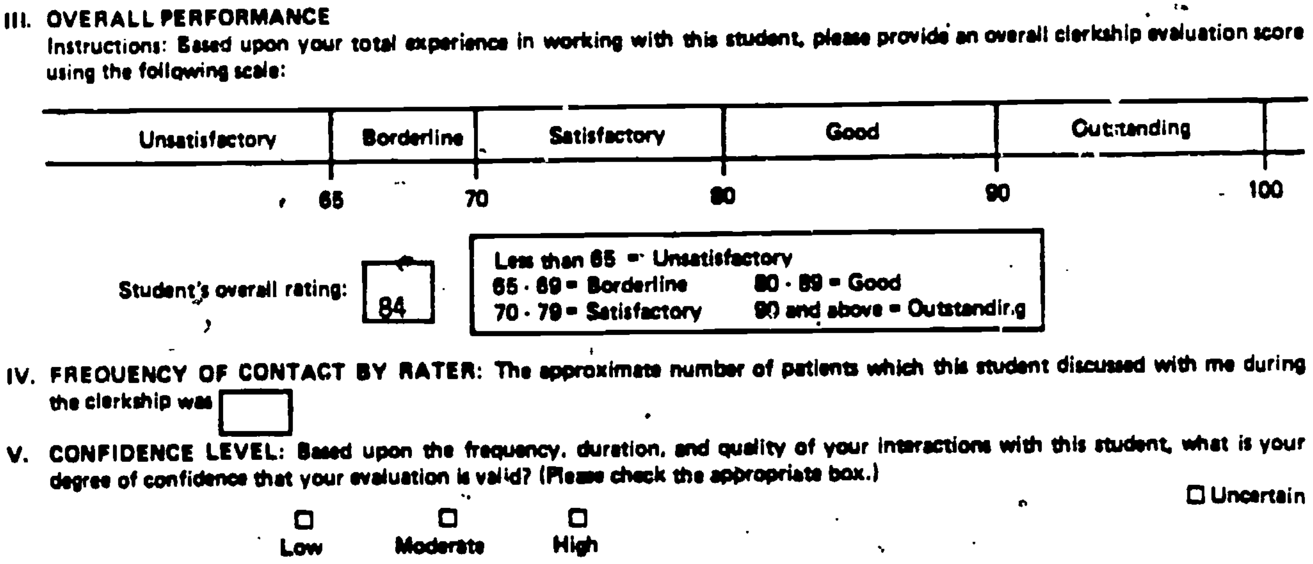
A first step in modifying rater behavior would be to ascertain whether individual raters know when their ratings are an accurate assessment of the ratee. Metacognition refers to knowing about one's own mental processes in order to perform intellectual tasks better (Flavell, 1979). Performance raters should have some sense of their accuracy in judging each ratee. If they are highly confident in their assessment of certain ratees, those ratings should be more accurate than other ratings. At some low threshold of confidence, a rater might disqualify him/herself from judging a particular ratee.

METHODS

Subjects in the study were 16 Family Practice faculty who completed 353 ratings of 59 senior medical students in a Primary Care Clerkship. The clerkship is a six week experience in treating ambulatory patients. Students are assigned to various clinics then rated at the end of the clerkship by 7 to 10 faculty and chief residents who worked with them. The rating form has been in use by the department for four years. It consists of five parts: 1. Attainment of Learning Objectives (6 scales), 2. Progress during the Clerkship (1 scale), 3. Overall Performance (1 scale), 4. Frequency of Contact (# of patients discussed) and 5. Confidence Level (low, moderate, & high). Figure 1 shows the three sections of the form used in this study.

Figure 1

Performance Rating Form



III. OVERALL PERFORMANCE

Instructions: Based upon your total experience in working with this student, please provide an overall clerkship evaluation score using the following scale:

Unsatisfactory	Borderline	Satisfactory	Good	Outstanding
65	70	80	90	100

Student's overall rating: 84

Less than 65 = Unsatisfactory
65 - 69 = Borderline   80 - 89 = Good
70 - 79 = Satisfactory   90 and above = Outstanding

IV. FREQUENCY OF CONTACT BY RATER: The approximate number of patients which this student discussed with me during the clerkship was □

V. CONFIDENCE LEVEL: Based upon the frequency, duration, and quality of your interactions with this student, what is your degree of confidence that your evaluation is valid? (Please check the appropriate box.)

□ Low   □ Moderate   □ High   □ Uncertain

Data analyses addressed the question. Is rater accuracy on each student significantly related to level of contac: or rater confidence? Rater accuracy was calculated for each of the 353 ratings. Two steps were required: 1. calculate a mean Overall Performance score (x_m) across raters for each of the 59 students, 2. subtract each individual rating (x_i) from its related mean ($s = x_m - x_i$). A positive score indicates a stringent rater. The accuracy score incorporates random error from both the original rating (x_i) and the individual student mean score (x_m), therefore the precision of each accuracy score is low. The independent variables were rater-reported level of contact with a student and confidence in the rating assigned. Level of contact (# of patients discussed with the student) was categorized into three groups of approximately equal size: 1–6 patients, 7–11 patients and 12–30 patients. The low and moderate rater confidence categories (see Figure 1) were combined due to the small number of low confidence ratings. A two factor analysis of variance (SPSS, 1983) was used to explore the relationships among accuracy scores, level of contact and rater confidence in the score assigned.

RESULTS and DISCUSSION

Table 1 displays mean accuracy scores and sample sizes for each of the six cells in the independent variable matrix.

Table 1

Rater Accuracy Mean Scores and Sample Size

	1-6 pts. discussed	7-11 pts. discussed	≥12 pts. discussed	Total
Low/Moderate Confidence	2.30 (107)	-.63 (68)	-1.16 (50)	.65 (225)
High Confidence	1.87 (20)	-1.01 (41)	-2.10 (67)	-1.13 (128)
Total	2.24 (127)	-.77 (109)	-1.70 (117)	

Table 2 is an ANOVA summary table for the two factor analysis of variance with Level of Contact and Rater Confidence as the independent variables. There were statistically significant differences in accuracy of ratings among the three levels of student contact. Scheffe's post hoc procedure indicated that the mean accuracy score for ratings with 1-6 patients discussed was lower than the other two scores. Differences in accuracy between low/moderate and high confidence ratings were not significant.

Table 2

ANOVA Summary Table

Source	SS	df	MS	F	Sig.
Combined main effect	1064.03	3	354.68	16.17	<.001
Level of contact	806.21	2	403.10	18.38	<.001
Rater confidence	27.18	1	27.18	1.24	.266
Interaction	4.832	2	2.42	0.11	.896
Residual	7610.02	347	21.93		

The significant differences in accuracy scores at three levels of student contact can readily be seen among the mean scores at the bottom of Table 1. Note that accuracy did not improve with increased rater exposure to the student. Instead, it appeared that in early contact with a student (1 to 6 patients discussed), raters gave overly stringent scores. They were most accurate when 7 to 11 patients were discussed and they became overly lenient with 12 or more patients discussed. Perhaps with extended interactions, a mentor/mentee relationship formed resulting in inflated ratings. The same statistical relationship between accuracy scores and level of student contact occured when the number of patients discussed was categorized into four levels (1-5, 6-9, 10-14 and >14) and tested by ANOVA. In summary, performance ratings in this setting were most accurate when the rater discussed 7 to 11 patients with the student.

The difference between accuracy scores of low/moderate vs. high confidence ratings was not statistically significant. Of 37 low confidence ratings, 32 were associated with low student contact. This finding was predictable in that low contact with a student should reduce rater confidence. On the other hand, 20 high confidence ratings were based upon low student contact, with 14 of these ratings from one rater. A follow-up frequency distribution analysis of individual rater confidence levels identified 5 raters with a large percentage of low confidence ratings and 4 raters with unusually high confidence. Perhaps rater confidence reflects individual personalities or cognitive style and is at best a very crude measure of metacognition. Alternatively, one might conclude that rater metacognition is not related to the accuracy of a rating.

A better understanding of how raters observe, encode, store, retrieve and record information may require studies where individual raters are the unit of analysis. Raters in this study produced significantly different individual mean scores (79.04 to 88.59; grand mean = 83.75). They also exhibited different standard deviations (2.48 to 7.72; overall = 5.97). By contrast, the large residual sum of squares in Table 2 indicates that level of contact and rater confidence were not very powerful in accounting for the variance in these data. Reports of long term consistent differences in rater leniency error were cited in the Introduction. Making extreme judgements as an individual rater trait has also been reported; however, it was not generally related to rater personality (Warr & Coffman, 1970). Research which focuses on the judgement processes of individual raters is more likely to yield new insights than a focus on general variables such as metacognition. Construct psychology (Kelly, 1955) would predict that raters use a small number of basic dimensions (constructs) to appraise ratees. If one could identify individual rater constructs related to this setting, perhaps they would provide an insight into the apparent differences among raters in scores assigned. If the psychological processes used by raters were better understood, improved methods for developing rating criteria and training raters could be developed.

CONCLUSIONS

Metacognition of raters, as measured by their confidence in the validity of a rating, was not related to the accuracy of that rating. Level of rater-ratee contact was related to accuracy, with the most accurate ratings based upon 7 to 11 patients discussed. Low levels of student contact were related to overly stringent ratings and high levels were associated with lenient ratings. Individual raters differed with regard to leniency of scores, the tendency to make extreme judgements and confidence in each rating. Future performance ratings research should use individual raters as the unit of analysis to better understand the psychological processes used in making judgements.

Bibliography

1. Cason, G.J. & Cason, C.L. A deterministic theory of clinical performance rating: promising early results. Eval. & the Health Prof., 1984, 7(2), 221-227.

2. Feldman, J.M. Beyond attribution theory: cognitive processes in performance appraisal. J. of App. Psy., 1981, 66, 127-148.

3. Flavell, J.H. Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. Am. Psy., 1979, 34(10), 906–911.

4. Kelly, G.A. The Psychology of Personal Constructs, Vol. 1 and 2. Norton, New York, 1955.

5. Landy, F.J. & Farr, J.L. Performance rating. Psy. Bull., 1980, 87(1), 72–107.

6. Littlefield, J.H., Ellis, R.E., Cohen, P.A. & Herbert, R.J. Leniency and score distribution differences among clinical raters. Proceedings of the Twenty-third Annual Conference on Research in Medical Education, Association of American Medical Colleges, 1984.

7. Marienfeld, R.D. & Reid, J.C. Six-year documentation of the easy grader in the medical clerkship setting. J. of Med. Educ., 1984, 59(7), 589–591.

8. SPSS Inc., SPSS/Pro: SPSS for the DBC Professional 350, McGraw-Hill Book Company, New York, 1984.

9. Warr, P.B. & Coffman, T.L. Personality, involvement and extremity of judgement. Br. J. of Soc. & Cl. Psy., 1970, 9, 108-121.