DOCUMENT RESUME

ED 271 483 TM 860 305

AUTHOR Littlefield, John H.; Troendle, G. Roger
TITLE Rating Format Effects on Rater Agreement and Reliability.
PUB DATE Apr 86
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cognitive Processes; *Dental Evaluation; *Dental Schools; Evaluation Criteria; Higher Education; *Interrater Reliability; Judges; Measurement Techniques; *Medical School Faculty; *Rating Scales; Stimuli

ABSTRACT

This study compares intra- and inter-rater agreement and reliability when using three different rating form formats to assess the same stimuli. One format requests assessment by marking detailed criteria without an overall judgement; the second format requests only an overall judgement without the use of detailed criteria; and the third format combines detailed criteria with an overall judgement. Results are interpreted from a cognitive processing theoretical framework. Subjects were five full-time and three part-time dental faculty members. The experimental task was to evaluate five crown preparations during six trials using each of three different rating forms, but raters were not informed they were reevaluating the same teeth. Raters were assigned code numbers to maintain anonymity, and teeth were identified only by code numbers. Data analysis was based upon ratings of five teeth from trials one through six; the trials were six weeks apart. Inter-rater agreement among the eight raters was distressingly low, but was in the range of one previous report. The study suggests that the traditional practice of scoring performance ratings by summary across multiple criteria may reduce intra-rater reliability. Rating forms which are structured to parallel rater cognitive processes may result in more reproducible scores than traditional summation scoring methods. (LMO)

Rating Format Effects on Rater Agreement and Reliability

John H. Littlefield, Ph.D. and G. Roger Troendle, D.D.S., M.S.

University of Texas Health Science Center

at San Antonio

Paper Presented at the Annual Meeting of the American Educational Research Association, San Francisco, California April, 1986

Rating Format Effects on Rater Agreement and Reliability

Background

Summing across multiple items to yield a single total score is the traditional scoring method on rating forms used in education and industry. This practice is based on psychometric theory which recognizes that individual items have considerable specificity and measurement error (Nunnally, 1978). This scoring method may not be appropriate for performance ratings because a third party, the rater, produces the scored responses for the individual being assessed. The rater's information processing serves as a *cognitive filter* of the measurement data (Landy & Farr, 1980). If we hope to increase the validity of performance ratings, we must learn more about how raters observe, encode, store and retrieve information marked on rating forms.

In describing the cognitive process of performance ratings, Feldman (1981) hypothesizes that raters can attend to a particular stimulus configuration without conscious monitoring. He suggests that stimuli are categorized into *fuzzy sets* which are not defined by necessary and sufficient sets of attributes. A given stimulus (performance) is categorized based upon the extent to which it overlaps features of a rater's *category prototype* (e.g., young ambitious energetic employee). If a stimulus does not automatically fit a *category prototype*, a consciously controlled process will supersede the automatic process. Both the automatic and the consciously-controlled processes are based upon a prototype-matching operation.

Human judgmental heuristics and knowledge structures (Nisbett & Ross, 1980) undoubtedly affect the cognitive process of performance appraisal. The perceiver is not a dutiful clerk who passively registers items of information. Instead, human perceivers actively interpret incoming perceptual data and form inferences about associations and causal relations. Faculty who rate students are experts in the tasks to be evaluated. As experts, faculty have complex schematic cognitive structures which provide an interpretive framework for making judgments about student performance. These cognitive structures may not directly correspond to the detailed rating criteria listed on a rating form.

Given that faculty who rate students are experts at the tasks to be judged, it seems unlikely that they strictly attend to the directions of a traditional performance rating form (*i.e., judge criterion 1, judge criterion 2, ..., sum the criterion scores*). Instead, it seems more likely that they match the stimulus performance against their own preconceived *category prototypes* and then make a global judgment. Marking detailed criteria may occur in conjunction with the global judgment, but would not necessarily precede it as implied by summing the criterion scores to yield a total score.

In summary, traditional performance rating forms are not structured to parallel the hypothetical cognitive processes used by experts to make judgments (e.g., raters are not asked to make an overall judgment about the performance). Instead, the rating form is constructed by logically analyzing

components may be important in helping a learner to analyze the multiple steps in a task, but they are not necessarily useful criteria for expert raters evaluating a stimulus performance.

This study compares intra- and inter-rater agreement and reliability when using three different rating form formats to assess the same stimuli. One format requests assessment by marking detailed criteria without an overall judgment. The second format requests only an overall judgment without the use of detailed criteria. The third format combines detailed criteria with an overall judgment. Results are interpreted from a cognitive processing theoretical framework.

Methods

Subjects were five full-time and three part-time dental faculty members. They ranged in age from 28 to 60 years. All subjects had 2 or more years of clinical teaching experience in the Division of Crown and Bridge and had participated in construction of the detailed rating criteria used in this study.

The rating task in this experiment is a routine part of the subjects' daily responsibilities. The experimental task was to evaluate five 3/4 crown preparations twice using each of three different rating forms:

1. *Form CrC* - a 19 item criterion checklist in which raters marked each criterion on a 3 category scale (acceptable, needs improvement, or unsatisfactory). A single composite score was calculated ex post facto by summing the marks on the 19 individual criteria (typical checklist).
2. *Form GJ* - a global judgment on a 5 point scale (0-4) with no detailed criteria.
3. *Form Com* - a combination of the 19 item criterion checklist (Form CrC) plus global judgment (Form GJ). The rater marked the individual criteria and also made a global judgment on a 4 point scale (0, 2, 3, 4).

Appendix 1 is a sample of Form Com. Note that the grading code allows 3 "I" ratings (improvement needed) to receive a grade of "2" while 4 "I" ratings results in a grade of "0" (failure). The omission of a "1" in the grade code reflects an evaluation philosophy which requires a satisfactory performance level to attain *clinical acceptability*. The occurence of 1 "U" rating or 4 "I" ratings results in a judgment of "failure" (0) for the crown preparation.

Table 1 summarizes the design of the study.

Table 1 - Design of the Study

Rating Form	Criterion Checklist (CrC)		Global Judgment (GJ)		Combination (Com)	
Trial Number	1	2	3	4	5	6
# of Raters	8	8	8	8	8	8
# of Teeth	15	5	5	5	5	5

Data collection procedures were described in detail by Troendle (1983). Raters were assigned code numbers to maintain anonymity. Fifteen crown preparations (teeth) were evaluated during trial one as shown in Table 1. For trials 2 through 6, five teeth were selected based upon the trial 1 ratings: a. two teeth that were easy to evaluate (high inter-rater agreement), b. two that were difficult to evaluate (low agreement), and c. one tooth that was of intermediate difficulty. Raters were not informed that they were re-evaluating the same five teeth. Teeth were identified only by code numbers and at least six weeks intervened between each trial session. Data analysis was based upon ratings of five teeth from trials 1 through 6.

Three types of *scores* were available for analysis:

1. Ratings on detailed criteria (Forms CrC and Com only)
2. Summated scores calculated by assigning 2, 1, or 0 to each A, I, or U rating on the detailed criteria then summing across the 19 detailed criteria (Forms CrC and Com only).
3. Competency-based scores using the 4,3,2,0 grading code shown in Appendix 1. Subjects provided this score when using Forms Com and GJ while the authors calculated it for Form CrC.

The term competency-based was used to signify the discontinuous score scale inherent in the *clinical acceptability* evaluation philosophy described above. Scores from Form GJ were classified as competency-based because this grading procedure is routinely used in the Dental School and was familiar to the raters. Table 2 summarizes the types of scores available for each rating format.

Table 2 - Three Types of Rating Data

	Form CrC	Form GJ	Form Com
Ratings of Detailed Criteria	Yes	No	Yes
Summated Scores	Yes	No	Yes
Competency-based Scores	Yes	Yes	Yes

Rating data were analyzed to answer four questions related to intra- and inter-rater agreement and reliability. Agreement was analyzed only on the ratings of detailed criteria (see Table 2) and was defined as identical ratings on a criterion. Two rater *agreement* questions were addressed:

1. Intra-rater agreement on the detailed criteria?
2. Inter-rater agreement on the detailed criteria?

Intra- and inter-rater agreement on the detailed criteria were assessed using a *tau* coefficient suggested by Tinsley and Weiss (1975). It is a chi square test to ensure that observed agreement exceeds chance levels followed by calculation of percent agreement adjusted down for chance agreements.

Reliability was defined as the degree to which the overall scores (both summated and competency-based in Table 2) are proportional when expressed as deviations from the judges' mean score. Two rater *reliability* questions were addressed:

1. Intra-rater reliability on the overall scores?
2. Inter-rater reliability on the overall scores?

Intra-rater reliability on the overall scores was assessed using a Pearson product-moment correlation coefficient and inter-rater reliability was assessed using an intraclass correlation coefficient (Finn, 1970). Statistical significance of differences among correlation coefficients was assessed using a multiple comparison test suggested by Marasculio (1966).

Results

Table 3 presents a summary of the data analysis results.

Table 3 - Data Analysis Results

	Form CrC	Form GJ	Form Com
Agreement on Detailed Criteria			
1. Intra-rater	75%	---	86%
2. Inter-rater	9% ††	---	36%
Reliability of Summated Scores			
1 Intra-rater	.59	---	.83 †
2. Inter-rater	.50	---	.56
Reliability of Comp.-based scores			
1 Intra-rater	.55	.73	.88 †
2. Inter-rater	.51	.63	.55

† - brackets denote significant differences (p<.05)
†† - does not exceed chance agreement

Inter-rater agreement on Form CrC (9%) did not exceed chance level. Intra-rater reliability coefficients for Forms CrC, GJ, and Com using both competency-based and summated scores were significantly different (p<.05). Form CrC coefficients differed from Form Com for both types of scores as shown by the brackets in Table 1. Inter-rater reliability coefficients associated with each Form were not statistically different.

Discussion

The intra-rater agreement levels in Table 3 parallel previous reports (Houpt & Kress, 1973). Inter-rater agreement among the 8 raters is distressingly low (i.e., did not exceed chance levels), but is in the general range of one previous report (Natkin & Guild, 1973). When inter-rater

agreement and intra-rater reliability results are viewed together, this study suggests that scores are more reproducible when the rating format requests a global judgment in conjunction with marking detailed criteria (Form Com). This is not the scoring procedure traditionally used with performance ratings (Form CrC).

Form CrC is typical of checklists which assume that expert raters evaluate a performance by assessing each detailed criterion individually while marking the corresponding blanks on the form. Summing the marks to get a single composite score could be delegated to a computer. However, recent reports on the role of prior knowledge in comprehension of medical information showed that experts made more inferences on high relevance information than either novices or intermediates (Patel, et. al., 1984). The raters in this study were experts at the task of preparing teeth for crown restorations. In judging a crown preparation, it seems likely that they would form an overall judgment based upon high relevance information in conjunc- tion with assessing each of the detailed criteria. Cognitive processes such as *confirmationist orientation* (Cooper, 1981) would influence a rater toward marking the detailed criteria to conform with his/her initial overall judgment. In summary, the structure of Form Com encourages an overall judgment and therefore parallels the sequence of rater cognitive processes more closely than form CrC. The *parallel structure* between Form Com and rater cognitive processes may have resulted in the improved intra-rater reliability for Form Com under both summated and competency-based scoring methods.

Mackenzie et. al., (1982) describe the results of a detailed investigation of rater error using dental performance checklists similar to those in this study. They conclude that, "... clearly defined unambiguous checkpoints are probably the most important factors in producing reasonable agreement among evaluators". Mackenzie et. al. also note that a criterion used to judge dental products is often based upon an opinion without validating whether clinical usefulness is impaired when that criterion is not satisfactorily achieved. The results of this study can be combined with the Mackenzie et. al. conclusions to suggest guidelines for developing rating form criteria and scoring procedures:

1. Develop clearly defined unambiguous checkpoints.
2. Use only criteria which can be shown to impair clinical usefulness when not satisfactorily achieved.
3. Use a scoring procedure which facilitates the ability of experts to distinguish between good and poor performances.
4. Use global judgment scoring in combination with marking detailed criteria (Form Com) instead of summing across the individual criteria.

In summary, the rating form should reflect cognitive processes used by experts judging the procedure rather than trying to *train* experts to use the form consistently.

Construct psychology (Fransella & Bannister, 1977) offers a technology for identifying the *actual* criteria used by experts to make judgments. The basic approach is to show expert raters examples of good and bad performances then ask them how various pairs differ. The final results are constructs used

by experts to distinguish among performances rather than logical steps used to teach the skill to a novice. The concept underlying identifying rater constructs is quite similar to the *retranslation* technique originally proposed to develop Behaviorally Anchored Rating Scales (Smith & Kendall, 1963), but it is less time consuming.

This study is marred by at least four weaknesses. First, the raters knew the data were for research purposes. Landy and Farr (1980) noted that ratings for administrative purposes will be more lenient than those for research purposes. Raters in this study may have performed differently if the scores were to be used to determine student grades. A second weakness is the failure to use a randomized block design. One could argue that the raters *learned* the teeth in rating them six times. The teeth were numbered with different ink and tape for each trial and stored loosely in a box. Posthoc conversations with the raters did not indicate that they recognized the same crown preparations were being used repeatedly. A third problem is the use of parametric statistics with a discontinuous competency-based scoring scale (0,2,3,4). This may have affected the size of the competency-based reliability coefficients; however, methodological studies of factor analyses with numerical scales that are not equal interval suggest that the correlation coefficients will not be substantially affected (Baggaley & Hull, 1983). In addition, the summated score reliability coefficients also support the superiority of Form Com (see Table 3). Finally, the fourth weakness is a failure to find significant differences among the inter-rater reliability coefficients. DiStefano (1981) has shown that intra-class correlation coefficients have a large sampling error when based upon relatively small sample sizes such as these (n=40 scores).

Conclusions

This study suggests that the traditional practice of scoring performance ratings by summing across multiple criteria may reduce intra-rater reliability. The results are consistent with a cognitive process of prototype matching in which the overall judgment made by an expert rater is an important part of the performance evaluation process. Rating forms which are structured to parallel rater cognitive processes (i.e., request experts to make an overall judgment somewhat independent of the detailed criteria) may result in more reproducible scores than traditional summation scoring methods. Detailed criteria are important to document the rationale which supports a rater's global judgment; however, it is not likely that criteria on a rating form mechanically structure the rater's judgment process as implied in traditional performance checklist directions and scoring. In short, the whole score may be different than the sum of the detailed criteria.

Problems with low inter-rater agreement in marking detailed criteria on rating forms may be due to the use of inappropriate criteria. Logical steps used to teach students to perform a procedure may not be helpful to experts in making consistent discriminative judgments. Techniques from construct psychology are recommended to help identify criteria which experts use to distinguish between good and poor performance of a particular task.

Bibliography

Baggaley, A.R. & Hull, A.L. The effect of nonlinear transformations on a likert scale. *Eval. & Hlth. Prof.*, 1983, 6(4): 483-491.

DiStefano, J.A. Sampling Error of Estimates of a Multifacet Generalizability Coefficient, Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, California, April, 1981.

Feldman, J. Beyond attribution theory: cognitive processes in performance appraisal. *J. Applied Psy.*, 1981, 66: 127-148.

Finn, R.H. A note on estimating the reliability of categorical data. *Educ. & Psy. Meas.*, 1970, 30:71-76.

Fransella, F. & Bannister, D. *A manual for repertory grid technique*. New York: Academic Press, 1977.

Houpt, M.I. & Kress, G. Accuracy of measurement of clinical performance in dentistry. *J. Dent. Educ.*, 1973, 37:34-46.

Landy, F.J. & Farr, J.L. Performance rating. *Psy. Bull.*, 1980, 87:72-107.

Marasculio, L.A. Large-sample multiple comparisons. *Psy. Bull.*, 1966, 65(2): 280-290.

Mackenzie, R., Antonson, D., Weldy, P., Welsch, B. & Simpson, W. Analysis of disagreement in the evaluation of clinical products. *J. Dent. Ed.*, 1982, 46(5): 284-289.

Natkin, E. & Guild, R.E. Evaluation of preclinical laboratory performance: A systematic study. *J. Dent. Educ.*, 1973, 37:152-161.

Nisbett, R. & Ross, L. *Human Inference: strategies and short-comings of social judgment.* Englewood Cliffs, N.J., Prentice-Hall Inc. 1980, 17-62.

Nunnally, J. C. *Psychometric Theory*, New York: McGraw-Hill, 1978, p. 84.

Patel, V., HoPingKong, H., Mark, V. Role of prior knowlede in comprehension of medical information by medical students and physicians. *Research in Medical Education: 1984, Proceedings of the Twenty-third Annual Conference on Research in Medical Education*, Association of American Medical Colleges, 1984.

Smith, P.C. & Kendall, L.M. Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. *J. App. Psy.*, 1963, 47(2): 149:155.

Tinsley, H.E. & Weiss, D.J. Interrater reliability and agreement of subjective judgments. *J. Coun. Psy.*, 1975, 22(4):358-376.

Troendle, G.R. The effects of three rating forms on intrarater and interrater agreement and reliability in the rating of 3/4 crown preparations. Unpublished thesis, University of Texas Graduate School of Biomedical Sciences, San Antonio, Texas, 1983.

APPENDIX 1

FORM CR-X: THE CRITERION-REFERENCED FORM WITH A COMPOSITE SCORE FOR TRIALS 5 AND 6

Code # _____

Tooth # _____

Date _____

Grading Code
0 - 1 I = 4
2 I = 3
3 I = 2
4 I = 0
U = 0

PREPARATION (3/4 Crown)

A	I	U	
☐	☐	☐	Axial reduction (over/under reduced)
☐	☐	☐	Occlusal reduction (over/under reduced)
☐	☐	☐	No sharp line angles present
☐	☐	☐	Two-plane reduction utilization
☐	☐	☐	Margin location
☐	☐	☐	Margin smoothness, continuity
☐	☐	☐	Margin type
☐	☐	☐	Occlusal convergence 2° - 5°
☐	☐	☐	Occlusal convergence less than 15°
☐	☐	☐	No undercuts present
☐	☐	☐	Position of proximal boxes slightly buccal to the middle of the tooth
☐	☐	☐	Proximal boxes in same line of draw
☐	☐	☐	Line of draw of boxes with the rest of the prep
☐	☐	☐	2° - 5° occlusal divergence of proximal boxes
☐	☐	☐	Blending of occlusal groove with proximal boxes
☐	☐	☐	Length of box
☐	☐	☐	Depth of box
☐	☐	☐	Margin below gingival floor of box
☐	☐	☐	Other* _____

GRADE