DOCUMENT RESUME

ED 104 942 TM 004 400

AUTHOR	Cohen, Stuart J.; Bengston, John K.
TITLE	A Psychophysical Investigation of Factors Affecting Teacher-Observers' Judgment.
SPONS AGENCY	Toledo Univ., Ohio. Graduate School.
PUB DATE	Apr 75
NOTE	15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 5, 1975) (Abridged)
EDRS PRICE	MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS	Evaluation Methods; *Observation; Pictorial Stimuli; Psychophysiology; *Rating Scales; Scoring Formulas; Simulation; *Social Reinforcement; Student Attitudes; *Student Reaction; Teacher Behavior; Teacher Evaluation; *Teacher Rating; Test Reliability; Video Tape Recordings

ABSTRACT

One hundred twenty-eight observers randomly assigned to 16 treatment conditions in a modified Latin square design, viewed three videotapes of simulated classrooms in which teacher behavior was controlled (paralleling psychophysical procedures) to fit unambiguously into specific categories on ratings of frequency and variety of social reinforcement. True behavior deimension scores, person and performance consistency, and inferential level of coding forms were manipulated to determine their effects on frequency and variety ratings and six semantic differential items. Results of multivariate analyses of variance indicated that stimulus variables and observation system characteristics significanlty affected the mean, variance, and accuracy of observers' judgments. (Author)

* A PSYCHOPHYSICAL INVESTIGATION OF FACTORS AFFECTING TEACHER-OBSERVERS' JUDGMENT[1]

Stuart J. Cohen, University of Toledo
John K. Bengston, University of Toledo[2]

Field research on the effects of teacher behavior on learner outcomes, competency-based assessment of teacher interns, as well as the process evaluation of in-service teachers use observer judgments as their primary data source. In such cases the usual index of reliability discussed in inter-observer agreement, in spite of repeated warning sound by Medley and Mitzel (1963), McGaw, Wardrop, and Bunda (1972), and Frick and Semmel (1974). These authors note that observer agreement is a component of estimates of reliabilities of observational records but by itself is an inadequate estimate. Yet, as a contributor to the reliability of observational records, observer disagreement cannot be disregarded either by those constructing observational systems or those training observers for classroom research.

Frick and Semmel (1974) state that "minimal observer disagreement is a necessary but insufficient condition for high reliability coefficients,since there are other components of the generic error variance which are theoretically independent from observer error variance" (p. 3). Among some of these contributers are instability of the behaviors under scrutiny and poorly designed observational systems. Both the stability of the underlying behavior and the nature of the observational system can affect observer agreement. It might even

[1] Abridged version of a paper presented at the annual convention of the American Educational Research Association, Washington, D.C. April, 1975.

[2] Now at the University of Florida, Gainesville, Florida.

* This research was funded by a grant from the Graduate School of the University of Toledo.

be the case that low observer agreement is less a function of observer error than of real ambiguity in the world and hence a positive index of the accuracy of a set of judgments.

If the "true scores" for the behaviors being observed and recorded were known, the stability of the behaviors and the accuracy of the observer judgments could be determined. Comparison to an "expert's" judgment would seem to provide a way out of this methodological mire where it not for the problem of finding some means of validating the accuracy of the expert. Clearly what is needed is an independent measure of the behavioral dimensions under scrutiny paralleling the physical measure of the stimulus available to the investigator conducting a psychophysical study of perception. Frick and Semmel (1974) suggested the creation of videotaped segments of a simulated teaching situation as a means of accomplishing this control. Editing and/or the use of prepared scripts would permit the removal of ambiguous instances of the critical dimensions, thereby providing "true scores" against which variance in observer judgments could be examined.

Rather than determine the virtues of a particular observational system or observer, the authors created segments of simulated classroom teaching behavior, to investigate factors that might have an effect on teacher observational in general. Frequency and variety of social reinforcement were selected as the behavioral dimensions for a number of reasons. First, there is copius documentation of the relationship between teacher reinforcement and student behavior (see, for example, Thoresen, 1972). Second, although the labels may vary, reinforcement categories appear within a number of different observation systems. Finally, by having judges rate variety as well as frequency, it was possible to evaluate the effects of the independent stimulus variables on judgments of differing conceptual complexity. Variety, a relational concept, is of course the more complex of the two.

The levels of these two dimensions were systematically manipulated to determine how true differences are reflected in the .ean, accuracy, and variance of their perceived levels. In addition, three other independent variables were investigated: person (making successive judgments of the same or different teachers); sequence of level of performance (judging the same or different levels of performance); and form (using a high or low inference coding form--the difference being in the specificity with which the critical dimensions are defined).

By scripting teacher behavior along the two critical dimensions, a control of the stimulus was achieved comparable to that obtained in psychophysical experiments. It was thus possible to ask questions about the functional relationship between the observer and the observed, without relying on the former to hazard a guess as to the tr_e value of the latter.

Methods

The two major dependent variables were ratings of frequency and variety of social reinforcement. Two forms of each scale were developed. The low inference forms contained six category levels, each with a label and a behavioral description (Appendices A and B). The high inference forms were identical to the low inference forms except that the behavioral descriptions were deleted (Appendix C).

Videotapes of simulated classroom situations were created to have "true scores" on the low inference forms by prompting teacher-action through cue cards to emit the exact number and variety of social reinforcers during their 11 minute drama lesson. A pool of both verbal and non-verbal reinforcing cues were developed and randomly assigned to each tape for the appropriate categories.

The order of these cues was also randomly determined. The teacher then had to find occasions in the spontaneous students' behavior to issue these reinforcers. The behavior of the two to four different high school students randomly assigned to participate in each lesson was not rehearsed, nor was the behavior of the teachers on the noncritical dimensions. During the videotaping, two observers watched for any ambiguity or errors of omission or commission which would require reshooting the entire lesson. A number of lessons were interrupted and reshot. At the conclusions of all videotaping, two observers again searched for any errors or ambiguities.

The four teacher-actors produced a total of nine different lessons. Each lesson had "true scores" on both frequency and variety of either three, four or five (categories C, D, and E respectively on the rating forms). The tapes were arranged in a modified Latin square depicted below. The capital letters stand for the teacher-actor involved, the Arabic numerals for the "true score" level of the tape, and the Roman numeral for the lesson for actor A, who has three tapes at the same criterion level.

Treatment Condition	Tape #1	Tape #2	Tape #3
1	A-3	A-4-I	A-5
2	A-4-I	A-5	A-3
3	A-5	A-3	A-4-I
4	B-3	C-4	A-5
5	C-4	D-5	A-3
6	D-5	B-3	A-4-I
7	A-4-II	A-4-III	A-4-I
8	D-4	C-4	A-4-I

The use of either the high or low inference rating form doubled the design to produce sixteen treatment conditions, to which 128 paid volunteer undergraduates who had never had classes with any of the teacher-actors were randomly assigned.

The study ran for six weeks. To guard against possible contamination across rating forms resulting from feedback students might give to their peers, students were randomly assigned to either the first or second three week time period. During the first three weeks all the high inferences subjects were run for all eight treatment conditions the sequence of which was randomly determined. This procedure assured randomization of treatment assignment and also guarded against contamination of the high inference form.

Each subject was seated at a desk in the experimental cubicle in front of a television monitor attached to a videotape casette playback deck. Each subject was given the appropriate one page instruction sheet for either the low (Appendix D) or high (Appendix E) inference rating forms and the forms themselves. After the S indicated he was ready to view tape #1, he was instructed to wear a head set which deleivered the audio. At the end of each trial, S completed the rating forms which were collected by E who then distributed new forms. After rating tape #3 for frequency and variety, Ss were given forms (Appendix F) for describing the third tape teacher or seven point sematic differential scales composed of bipolar adjectives. Most of the adjectives were selected because of their previous relationship with student achievement (Rosenshine and Furst, 1971). The order and position of the semantic adjectives was randomly determined to reduce the possibility of position or set bias. Upon completion o' this task, subjects were debriefed and asked not to discuss the nature of the study until after a given date.

Results and Conclusions

A number of different analyses were performed. Space in this discourse permits only a discussion of some of the major results. A 3X2X2 multivariate analysis of variance (MANOVA) was performed for level of reinforcement, same or different teacher throughout, and high or low inference coding form for rating of frequency and variety on the third trial tape. The overall MANOVA F for reinforcement levels was significant beyond .0001, as were the univariate F's for frequency and variety. The overall MANOVA F for coding form was also significant beyond ($p<.0001$). The probability associated with the univariate F for frequency was significant at .0001. There were no significant MANOVA interactions.

A 2 (consistent vs. different level for all trials) X 2 (person) X 2 (form) MANOVA produced significant overall effects for form only ($p<.0001$). This form effect was found in the univariate analysis for frequency ($p<.0001$). In addition to the two MANOVA's on the ratings of the third tape, there were univariate analyses for frequency and variety ratings on each of the first two trials. In all cases, differences in rating of both frequency and variety were significant well beyond .01. Thus, the judges ratings reflected the actual "true score" differences for the two critical dimensions. This provides additional credence for the psychophysical validity of the videotapes for the two dimensions manipulated. The significant differences between the forms apparently result from a combination of overstimation of the actual scores by users of the high inference forms, and underestimation of the actual scores by users of the low inference forms. To further investigate this phenomenon, each score for frequency and variety was subtracted from the "true score" for that cell, and MANOVA and univariate analyses were performed for the third tape ratings. There were no significant 2X2X2 MANOVA's. The only significant 3X2X2 MANOVA was for the form effect ($p<.02$). The low inference form proved more accurate on both ratings of frequency ($p<.06$) and variety ($p<.006$).

To assess the variability of observer judgments, each score was subtracted from its cell mean, and MANOVA and univariate analyses were then performed on the resulting absolute difference scores. Levene proposed such a procedure (Glass, 1966) for testing homogeneity of variance and asserted that these difference scores met the assumptions necessary for analysis of variance. The difference score results were strikingly similar to the accuracy analyses with an interesting reversal. Again there were no significant 2X2X2 MANOVA's. The only significant 3X2X2 MANOVA was for the form effect ($p<.007$). The low inference coding forms produced significantly less variance on ratings of both frequency ($p<.002$) and variety ($p<.05$).

The six semantic differential ratings of the teacher on the third tape were included with the variety and frequency ratings in MANOVA and univariate analyses. The 3X2X2 MANOVA revealed that teachers who had higher "true scores" on frequency and variety of social reinforcment were rated significantly higher ($p<.001$) on the dimensions of friendliness, acceptance, and sincerity. In addition, those teachers judged by raters using the high inference form, were rated significantly more friendly ($p<.02$).

Educational Importance of the Study

This study demonstrates how, by emulating the methodology of psychophysical experiments, observer judgments such as the level of performance of the teacher, seeing the same or different teachers, and the behavioral criteria of the rating form can be systematically investigated. This study documents how the inference level of the coding category can affect the accuracy and variability of judges' ratings of behaviors differing in complexity.

References

Frick, T. and Semmel, M.I. Observational records: Observer agreement and reliabilities. Center for Innovation in Teaching the Handicapped, Indiana University. Paper presented at the American Educational Research Association convention, April 1974.

Glass, G.V. Testing homogeneity of variances. American Educational Research Journal, 1966 3 (3), 187-190.

McGaw, B., Wardrop, J.L., and Bunda, M.A. Classroom observation schemes: Where are the errors? American Educational Research Journal, 1972, 9 (11), 13-27.

Medley, D.M. and Mitzel, H.E. Measuring classroom behavior by systematic observation. In N.L. Gage (ed.) Handbook cf research on teaching. Chicago, Ill.: Rand-McNally, 1963, 247-328.

Rosenshine, B. and Furst, N.F. Research on teacher performance criteria. In B.O. Smith (ed.) Research in teacher education: A symposium. Englewood Cliffs, N.J.: Prentice Hall, 1971, 37-72.

Thoresen, C.E. (ed.) Behavior modification in education, the seventy-second yearbook of the National Scoiety for the Study of Education. Chicago, Ill.: University of Chicago Press, 1973.

Appendix A

Date ____________

Rater's Name ____________________ Videotape Example # ______________

RATING FORM

Place an X in one blank only for each judgment.

JUDGMENT I: How frequently did the teacher give positive, social reinforcement to his students?

____A. Rarely (0-4 total instances)

____B. Seldom (5-9 total instances)

____C. Occasionally (10-14 total instances)

____D. Regularly (15-19 total instances)

____E. Often (20-24 total instances)

____F. Very Frequently (more than 25 total instances)

Note: For Judgment I if you observed a total of 10 positive, social reinforcers given by the teacher, you should place an X in the blank marked for "B. Seldom." If you observed 16 total instances, you should place an X in the blank for "D. Regularly."

Appendix B

Date ____________

Rater's Name ________________ Videotape Example # ____________

JUDGMENT II: Which of the following best describes the variety of positive social reinforcers given by the teacher?

_____ A. Extremely limited: All reinforcement given was within one of the three categories (positive verbal feedback, verbal praise, or non-verbal approval), and within that category there were no more than two different responses used.

_____ B. Very limited: All reinforcement given was within one category, but within that category there were more than two different responses used. OR, the teacher gave reinforcers from two of the categories, but did not use more than two different responses within each of those categores.

_____ C. Limited: All reinforcement given was within two categories, and the teacher gave more than two different responses in one category with no more than than two different responses in the other. OR, all three categories were respresented with the teacher giving no more than two different responses in each one.

_____ D. Somewhat varied: All reinforcement given was within two categories, and the teacher gave more than two different responses within both categories. OR, all three categories were represented with the teacher giving more than two different responses in one category, but only one or two different responses in the remaining two.

_____ E. Varied: All categories of reinforcement are used; in two categories the teacher gave more than two different responses and in one category the teacher gave only one or two different responses.

_____ F. Quite varied: All three categories of social reinforcement were used and in each category the teacher gave more than two different responses.

Note: For Judgment II your response should be based on both the category of social reinforcer (positive verbal feedback, verbal praise, or non-verbal approval) and the variety of social reinforcer within each category. It is important to distinguish whether the teacher repeats the same words or gestures within each category or whether he uses a number of different kinds of words or gestures. For example, a teacher who said "right" five times and said "O.K." six times is making a total of 11 positive verbal feedback statements. For Judgment I this teacher would be marked "C. Occasionally" but for Judgment II he would be marked " A. Extremely limited" since all his social reinforcers fell within the same category (positive verbal feedback) and he did not use more than two different responses within that category. Had the same teacher in the previous example also added one smile, one wink, one pat on the back, and said one "correct, " he would have made a total of 15 responses and would still be marked "C. Occasionally" for Judgment I. However, for Judgment II this teacher would now be marked "D. Somewhat varied" since all social reinforcement fell within two categories (positive verbal feedback and nonverbal approval) and there were more than two different responses within each category.

Appendix C

Date_______________

Rater's Name ________________________ Videotape Example # ________________

Indicate your rating of the frequency and variety of positive social reinforcements used by the teacher during the segment. Put an X in only one blank for each question.

ᵀ. Frequency of positive, social reinforcement

_____ A. Rarely

_____ B. Seldom

_____ C. Occasionally

_____ D. Regularly

_____ E. Often

_____ F. Very Frequently

II. V^ariety of pcsitive, social reinforcement

_____ A. Extremely limited

_____ B. Very Limited

_____ C. Limited

_____ D. Somewhat Varied

_____ E. Varied

_____ F. Quite Varied

Appendix D

Rater's Name ____________________ Date ______________

INSTRUCTION

This study is concerned with the a[illegible] education students to carefully observe and evaluate teaching performance. You will be viewing three, 11 minute, videotaped examples of instruction. Each example shows a drama teacher directing students in a scene from a play. Immediately FOLLOWING EACH example, you will be asked to make judgments about the amount and variety of positive, social reinforcement given by the drama teacher.

Please Note--The type of reinforcers to be watching for include:

A. Positive Verbal Feedback: Statements indicating correctness of response, for example: "O.K.," "That's right, Joe," "Correct, " "Fine," "exactly," etc.

B. Verbal Praise: In comparison to feedback, praise statements emphasize quality beyond correctness of response, for example: "Well done," "Good idea, Sue," "You're terrific!," "That's an interesting question," "Good," "Excellent," etc.

C. Non-verbal Approval: Clear and emphatic gestures of approval, for example: a broad smile, vigorous head nod, applause, a, pat, a hug, etc.

If you wish, you may keep notes on the rating form.

Appendix E

Rater's Name ________________________ Date ______________

INSTRUCTION

This study is concerned with the ability of education students to carefully observe and evaluate teaching performance. You will be viewing three, 11 minute, videotaped examples in instruction Each example shows a drama teacher directing students in a scene from a play. Immediately FOLLOW EACH example, you will be asked to make judgments about the amount and variety of positive, social reinforcement given by the drama teacher.

Please Note--in this study, positive, social reinforcement refers to the supportive things the teacher says or does as he interacts with students.

Read the rating form that has been provided. When you believe that you have an adequate understanding of the judgments you will be making, indicate that to the experimenter and he will run the first tape. If you wish, you may keep notes on the rating form.

Appendix F

Rater's Name ____________________________ Date ________________

Following is a list of paired adjectives with opposite meanings. They are located at the extreme points of a seven point scale representing the continuum of meaning for each pair. You are to rate for each pair of adjectives the behavior of the teacher in the <u>final</u> videotape examples only by checking the slot on the scale that you feel describes that behavior with regard to the adjective pair. For example, if "snow" were rated on a black/white scale, most people would probably check slot "7."

"snow"

black____/____/____/____/____/____/_X_white
1 2 3 4 5 6 7

"Coal" would probably receive a "1" and "twi-light" would be given a rating toward the middle ("4").

Teacher behavior for the final videotape example (example 3)

indifferent____/____/____/____/____/____/____friendly
1 2 3 4 5 6 7

accepting ____/____/____/____/____/____/____rejecting
1 2 3 4 5 6 7

clear ____/____/____/____/____/____/____confusing
1 2 3 4 5 6 7

dull ____/____/____/____/____/____/____stimulating
1 2 3 4 5 6 7

insincere ____/____/____/____/____/____/____sincere
1 2 3 4 5 6 7

static ____/____/____/____/____/____/____dynamic
1 2 3 4 5 6 7