DOCUMENT RESUME

ED 199 293 TM 810 244

AUTHOR Ory, John C.: Valois, Robert F.
TITLE The Influence of Negatively Worded Scale Items on Overall Ratings.
PUB DATE [80]
NOTE 13p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Course Evaluation: Higher Education: Item Banks: *Negative Forms (Language): Rating Scales: *Student Evaluation of Teacher Performance: Test Format: *Test Items
IDENTIFIERS *Instructor and Course Evaluation System

ABSTRACT

Two studies investigate whether the placement and/or wording (either positively or negatively) of diagnostic rating scale items influenced student responses to the global items in the evaluation of a course of instruction. The Instructor and Course Evaluation System (ICES) developed at the University of Illinois, Urbana-Champaign was used to conduct end-of-semester course evaluations. Thirty diagnostic items were selected from a catalog containing approximately 500 items. Half of the global items were about the course and half were about the instructor. Twenty of the 30 items were rewritten to create positively and negatively worded versions of each item. Three negative wording conditions were repeated on scales with the two global items appearing either before or after the 30 diagnostic items. In two studies, 455 undergraduates were randomly administered one of six evaluation forms. Results of a 2 x 3 analysis of variance with repeated measures indicated that the instructor was significantly higher rated than was the course. In neither study were the overall ratings of the instructor or course affected by the negative or positive wording of the diagnostic items. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED199293

The Influence of Negatively Worded Scale Items on Overall Ratings

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. C. Ory
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

John C. Ory
Robert F. Valois
University of Illinois

In his systematic study of response sets and their effects, Cronbach (1946, 1950) identified acquiescience as a response tendency to favor affirmative responses over negative responses. Couch and Keniston (1960) later called this tendency "yea- or nay- saying," wherein respondents consistently select in one direction, either positive or negative. Their belief was that some individuals have a general disposition on the positive/negative continuum regardless of the content of the items. Consequently, the responses of these same individuals may indicate something other than that which was intended to be measured.

To avoid response bias due to yea- or nay- saying, psychometricicms recommend counterbalancing the questions which were asked, so that a positive response to one question and a negative response to another both contribute towards increasing the score on the measure as a whole (Lemon, 1973; Likert, 1932; Edwards, 1957). Likert (1932) suggested that these "two kinds of statements ought to be distributed throughout the attitude test in a chance or haphazard manner (p. 91)."

For rating scales used to evaluate a new project, person or course of instruction, the above solution calls for the inclusion of both negatively and positively stated items about the object or person being evaluated. For example, course evaluation items should include items which make positive and negative statements about the course, such as:

TM 810 244

This course provided an opportunity to learn from other students. (positive)

Teaching methods used in this course were poorly chosen. (negative)

What has yet to be studied is the possible affect of negatively worded items on raters' evaluations. Do negatively worded items "encourage" a more critical evaluation than do positively worded items? Negatively worded items may highlight the negative aspects or faults of the object or person being evaluated, or may serve to unconsciously suggest to the rater particular problem areas anticipated by the evaluator. If so, rating scale evaluations may be affected as much by the wording of the items as by the quality of the object or person being evaluated.

The possible affect of item wording on overall ratings is particularly relevant to many of today's available student ratings instruments. Most of the available instruments include two kinds of scale items--global or generally stated items and diagnostic or specifically stated items. Global items measure student evaluations of general areas of instruction, while diagnostic items measure student judgments and observations of specific behaviors of the instructor, instructional techniques and detailed student outcomes. The following examples of each type of item are included on the Instructor and Course Evaluation System (ICES) developed at the University of Illinois, Urbana-Champaign (Note 1).

Global: Rate the course in general

____ ____ ____ ____ ____

Excellent Poor

Diagnostic: The instructor motivated me to do my best work.

____ ____ ____ ____ ____

Almost Always Almost Never

The working principle behind the ICES two-way classification of items (global and diagnostic) is that different types of items should be used for different purposes. The diagnostic items are best suited for the purpose of faculty improvement while the global items are most useful for providing summative information needed for personnel decisions (Brandenburg, Braskamp, and Ory, 1980). Faculty, therefore, select those diagnostic items they consider appropriate for their particular course. Each faculty questionnaire would also include three global items: Rate the course content, Rate the instructor, and Rate the course in general. Normative data are provided for the latter items only so that campus-wide comparisons can be made.

Unfortunately, little is known about the relationship between student responses to faculty-selected diagnostic items and global evaluation items. It has yet to be determined if the type of diagnostic item chosen by a faculty member can influence student responses on the global items. The purpose of these two studies was to investigate whether the placement and/or wording (either positively or negatively) of diagnostic rating scale items influenced student responses to the global items in the evaluation of a course of instruction.

Instrumentation

The ICES system was used to conduct the end-of-semester course evaluations. ICES is a cafeteria-type student rating system that permits each instructor to select diagnostic items from a catalog containing approximately 500 items. As was stated earlier, the first three items on all student questionnaires are global items. For purposes of this study, students responded to only the last two global items--Rate the instructor and Rate the course in general. Respondents indicated their rating on these two items on a 5-point scale, with anchor points of "poor" (=1) and "excellent" (=5). ICES questionnaires used in the

studies included thirty diagnostic items. Approximately half of the items were about the course and half were about the instructor. Twenty of the thirty diagnostic items were rewritten to create a positively and negatively worded version of each item. For example,

Positive version = Exams covered a reasonable amount of material.

Negative version = Exams covered an unreasonable amount of material.

In total, six evaluation forms were constructed containing 32 items each. The content and design of the six forms is explained in Figure 1. As illustrated, the three negative wording conditions (0/30, 10/30, 20/30) were repeated on scales with the two global items appearing either before or after the 30 diagnostic items. It was believed that if the wording of the diagnostic items was to influence student responses to the global items such effect may be more noticeable if the global items were presented after rather than before the diagnostic items.

	Proportion of negatively worded diagnostic items	Item Format
Form 1:	0/30	Global before diagnostic items
Form 2:	10/30	Global before diagnostic items
Form 3:	20/30	Global before diagnostic items
Form 4:	0/30	Diagnostic before global items
Form 5:	10/30	Diagnostic before global items
Form 6:	20/30	Diagnostic before global items

Figure 1: The content and design of the six evaluation forms.

Study One

Methods. During the last week of the 1980 Spring semester, 180 students enrolled in an undergraduate introductory health education course taught at a

Midwestern university were randomly administered one of the six evaluation forms. Thirty students completed each of the six forms.

Data Analysis. Differences in student responses to the two global items across the six evaluation forms were analyzed through a 2 x 3 analysis of variance (ANOVA) with repeated measures. The global assessments of course and instructor were repeated across the two placement conditions (global items before or after diagnostic items) and the three wording conditions (0, 10, or 20 of 30 items worded negatively). Resultant F-ratios were tested at a .05 level of significance.

Results. Global item means and standard deviations recorded on each of the six evaluation forms are presented in Table 1. Results of the ANOVA presented in Table 2 indicated that the instructor (4.71) was significantly ($p < .01$) higher rated than was the course (4.39). Also significant ($p < .01$) was the Type of global rating X Placement interaction. Inspection of the interaction cell means revealed that the overall ratings of the course were lower when the global items followed (4.22) the diagnostic items rather than preceded (4.52) them, whereas, the overall instructor ratings were approximately the same when presented either before (4.67) or after (4.74) the diagnostic items. While the lowest course and instructor overall ratings were obtained when 20 of the 30 diagnostic items were written negatively, there were no significant ($p < .14$) differences identified for either global rating across the three wording conditions.

Study Two

Results of Study One suggested that the placement more so than the wording of diagnostic scale items may influence student's responses to global items. However, limitations to the initial study prohibit a clear interpretation of

Table 1

Study One: Global Items Means and Standard Deviations Across the Six Evaluation Forms

	Placement Conditions							
	Before Diagnostic Items				After Diagnostic Items			
Wording Conditions	Instructor Ratings $\bar{X}$	SD	Course Ratings $\bar{X}$	SD	Instructor Ratings $\bar{X}$	SD	Course Ratings $\bar{X}$	SD
(0/30 negatives)	4.70	.47	4.73	.45	4.80	.48	4.23	.63
(10/30 negatives)	4.67	.84	4.47	.78	4.72	.68	4.28	.85
(20/30 negatives)	4.63	.67	4.37	.67	4.69	.60	4.17	.93

Table 2

Study One: ANOVA Summary Table

Source	df	SS	MS	F
Placement (P)	1	.81	.81	1.11
Wording (W)	1	.24	.12	.16
P x W	2	.54	.27	.37
Error	174	129.12	.73	
Type of Rating (T)	1	9.48	9.48	40.26*
T x P	1	2.90	2.90	12.30*
T x W	2	.94	.47	1.99
T x P x W	2	.67	.33	1.42
Error	174	41.45	.24	

*p < .01

the findings. First, ratings were collected in only one course taught by one instructor, therefore the generalizability of the findings to other courses and instructors is limited. Second, both the overall ratings of the course and of the instructor were quite high. Effects due to the negative wording of the diagnostic items may be more noticeable with less highly rated instructors and courses for which students have more to criticize. A second study was therefore conducted which was designed to remove these limitations.

Method. The six evaluation forms used in Study One were randomly administered during the last week of the 1980 Fall semester in 14 sections of an undergraduate introductory sex education and family life course taught by 7 instructors. Of the 275 students responding, approximately 45 responded to each of the six forms.

Data Analysis. The same 2 x 3 ANOVA with repeated measures used in the initial investigation was conducted.

Results. Means and standard deviations for the two global items recorded on each of the six evaluation forms are presented in Table 3. ANOVA results presented in Table 4 indicate significant ($p < .01$) differences between the overall ratings of the course (3.62) and instructor (4.13), with the latter ratings being higher. No significant ($p < .05$) differences were identified for either global rating across wording or placement conditions.

Discussion

In neither study were the overall ratings of the instructor or course affected by the negative or positive wording of the diagnostic items. In the first study, however, the placement of the diagnostic items influenced the global ratings of the course. The placement effects found in Study One indicated that students rated the course, but not the instructor, significantly lower when the global items were presented after rather than before the

Table 3

Study Two: Global Item Means and Standard Deviations Across the Six Evaluation Forms

	Placement Conditions							
	Before Diagnostic Items				After Diagnostic Items			
Wording Conditions	Instructor Ratings $\bar{X}$	SD	Course Ratings $\bar{X}$	SD	Instructor Ratings $\bar{X}$	SD	Course Ratings $\bar{X}$	SD
(0/30 negatives)	4.17	.86	3.54	1.07	4.23	.96	3.55	1.50
(10/30 negatives)	4.19	.83	3.77	.84	3.91	1.20	3.47	1.14
(20/30 negatives)	4.11	.90	3.73	.82	4.19	.97	3.71	1.04

Table 4

Study Two: ANOVA Summary Table

Source	df	SS	MS	F
Placement (P)	1	.71	.71	.51
Wording (W)	2	.93	.47	.34
P x W	2	3.23	1.62	1.16
Error	269	374.16	1.39	
Type of Rating (T)	1	35.24	35.24	48.43*
T x P	1	.10	.11	.15
T x W	2	1.49	.75	1.03
T x P x W	2	.03	.01	.02
Error	269	195.74	.73	

*$p < .01$

diagnostic items. An informal observation of students during the administration of the evaluation forms indicated that they responded to the last two global items after responding to the diagnostic items. Failure to find erasures of global item responses also indicated that students did not change their initial global item responses after completing the diagnostic items. These observations suggest that in responding to the diagnostic items first, the students may have used them as a type of "score card" for evaluating the overall quality of the instructor and course. The diagnostic items indicated instructional areas which needed to be considered in the global evaluations. Responding to the score card before making global assessments apparently lowered the students initial reactions to the course but not to the instructor. Possibly the students became more realistic in their assessments or were reminded of more weaknesses than strengths. Further research wherein students are asked to explain their responses to the evaluation forms is needed to find the correct explanation.

Why responding first to the diagnostic items altered the course ratings only is also not clear. Research has found that students require less prompting to discuss the strengths and weaknesses of an instructor than of a course (Braskamp, Ory, and Pieper, 1980) and are more consistent in their ratings of an instructor than of a course regardless of the evaluation method used; i.e., open-ended questionnaires, rating scales or group interviews (Ory, Braskamp, and Pieper, 1980). Students may therefore have a set opinion of the instructor already in mind and may not need the framework or prompting provided by the diagnostic items. On the other hand, the framework provided by the diagnostic items may help to narrow the range of areas needed to be considered when evaluating an entire course (i.e., exams, homework, lectures) and thus have greater impact on the ratings of the course.

More importantly than the reason(s) for the placement effects found in Study One, is the fact that such effects were not evident in the more externally valid second study. Neither wording nor placement effects were identified in the global ratings of 7 instructors teaching 14 course sections. Results of Study Two confirm the initial study's failure to find significant wording effects but fail to support the existence of the placement effects. Instead, the lack of findings in Study Two which would indicate possible sources of rating influence speaks well for the validity of student ratings. Students appear to have a general opinion about their instructor and course that is unaltered by the placement or wording of other scale items included on the evaluation form.

With the current increase in college and university use of student evaluations of instruction, the reliability and validity of student ratings is consistently being challenged. Numerous research studies (Aleamoni and Graham, 1974; Brandenburg, Slinde, and Batista, 1977; Frey, Leonard, and Beatty, 1975; Marsh, 1980; McKeachie and Lin, 1971) have investigated the extent to which extraneous variables (i.e., expected grade, class size, sex, timing of administration) bias the measurement of teacher and course quality. Results of these two studies add one (wording of diagnostic items) and possibly two (placement of diagnostic items) more extraneous variables to a growing list of factors which have little, if any, influence on global assessments.

Reference Notes

1. Illinois Course Evaluation System: Its rationale and description (ICES Newsletter No. 2). Urbana-Champaign: University of Illinois, Measurement and Research Division, Office of Instructional Resources, August (1977). (Mimeo).

References

Aleamoni, L. M. & Graham, M. H. The relationship between CEQ ratings and instructor's rank, class size, and course level. Journal of Educational Measurement, 1974, 11, 189-202.

Brandenburg, D. C., Braskamp, L. A. & Ory, J. C. Considerations for an evaluation program of instructional quality. CEDR Quarterly, 1979, 12, 8-13.

Brandenburg, D. C., Slinde, J. A. & Batista, E. E. Student ratings of instruction: Validity and normative interpretations. Journal of Research in Higher Education, 1977, 7, 67-78.

Braskamp, L. A., Ory, J. C., & Pieper, D. M. Written comments: Dimensions of instructional quality. Journal of Educational Psychology, in press.

Couch, A. & Keniston, K. Yea sayers and nay sayers: Agreeing response set as a personality variable. Journal of Abnormal and Social Psychology, 1960, 60, 151-174.

Cronbach, L. J. Response sets and test validity. Educational and Psychological Measurement, 1946, 6, 475-494.

Cronbach, L. J. Further evidence on response sets and test design. Educational and Psychological Measurement, 1950, 10, 3-31.

Edwards, A. L. Techniques of Attitude Scale Construction. New York: Appleton Century Crofts, 1957.

Frey, P. W., Leonard, D. W., & Beatty, W. W. Student ratings of instruction: Validation research. American Educational Research Journal, 1975, 12, 435-447.

Lemon, N. Attitudes and their Measurement. John Wiley & Sons, New York, 1973.

Likert, R. A technique for the measurement of attitudes. Archives of Psychology, 1932, 140, 44-53.

Marsh, H. W. The influence of student, course, and instructor characteristics on evaluations of university teaching. American Educational Research Journal, 1980, 17, 219-237.

McKeachie, W. J. & Lin, Y. Sex differences in student response to college teachers: Teacher warmth and teacher sex. American Educational Research Journal, 1971, 8, 22-226.

Ory, J. C., Braskamp, L. A., & Pieper, D. M. Congruency of student evaluative information collected by three methods. Journal of Educational Psychology, 1980, 72, 181-185.