ALTERNATIVE RESPONSE DEFINITIONS IN INSTRUCTIONAL RATING SCALES

Dissertation for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY
BARBARA HOUGHTON SHOWERS

1973

This is to certify that the

thesis entitled

ALTERNATIVE RESPONSE DEFINITIONS IN

INSTRUCTIONAL RATING SCALES

presented by

Barbara Houghton Showers

has been accepte
of the req

PhD de

Date August 9, 1973

ABSTRACT

ALTERNATIVE RESPONSE DEFINITIONS IN
INSTRUCTIONAL RATING SCALES

By

Barbara Houghton Showers

Over the years many efforts have been made to improve student ratings of teacher effectiveness. This study represents another such effort. It is concerned with the particular problem of the leniency bias shown by many students in rating their instructors. By leniency bias is meant the tendency of students to use only the two or three highest options in rating their instructors. The harmful effect of this bias is to reduce discrimination between instructors to the extent that small differences in mean ratings produce large differences in reported rankings. The idea which gave rise to the present study was that leniency bias could be reduced by changing the wording of the response options. It was hoped that a different wording would increase the range of options used by student raters and improve discrimination between instructors. The major reason for conducting the study was to improve an existing Likert-type student instructional rating scale. Since the content of the

scale was well established in its creation, the study was focussed on manipulating the response options to reduce the amount of lenient responding present with the existing scale. Two alternative response definitions were chosen to compare with the existing Likert format response definitions. The three response formats were, (1) fixed alternative Likert cues (SA-SD), (2) fixed alternative evaluative cues (superior-inferior), and (3) multiple choice short descriptive cues. A concurrent purpose of the study was to test two claims made in the literature concerning the bias-proneness of certain response definitions. The claims tested were:

a. Evaluative cues are more susceptible to bias than other cues.

b. Fixed response alternatives are more susceptible to bias than descriptive multiple choice alternatives.

It was hypothesized that the evaluative format would produce the most lenient responses, and the descriptive format the least lenient responses. It was also hypothesized that the least lenient response cue format would prove to have the greatest rater reliability, since a reduction in lenient responding would make possible improved discrimination between instructors.

To conduct the study, three instructional rating forms differing primarily in response cue format were developed and administered to random thirds of the classes of 23 instructors. Leniency bias was measured by finding

the closeness of each item mean to the midpoint of the rating scale. Since student ratings were overwhelmingly concentrated at the upper end of the scale, the format that gave the lowest mean was regarded as the least biased. The hypothesis of no differences in mean ratings (leniency bias) was tested with a two-way multivariate analysis of variance design, instructor by treatment, where the response cue formats were the three treatments and the 17 items were 17 dependent variables. Scheffe post hoc analyses tested alternate hypotheses that the evaluative format would produce the most lenient items, the Likert format the next most lenient, and the descriptive format the least lenient. The hypothesis of no differences in rater reliabilities was tested by comparing confidence intervals about the reliability estimate for each item. Non-overlapping confidence intervals would indicate significant differences in rater reliabilities.

The results of the study indicated that the evaluative format of the instructional rating scale was less prone to leniency bias than the other response formats. The evaluative format had less lenient means than either the descriptive or Likert formats for the majority of items. The Likert format, which was the format of the rating scale currently in use at the university, was found to be the most often prone to leniency bias.

Claims made in the literature concerning the proneness to bias of fixed alternative response formats in general, and evaluative formats in particular, were found not to hold with student ratings of instruction. Fixed alternative evaluative response cues were found to be the least susceptible to leniency bias in this study, while multiple choice descriptive response cues were found to be moderately susceptible, and fixed alternative Likert response cues most susceptible. Situational variables such as the purposes of the ratings and the experiences of the student raters were hypothesized to be somewhat responsible for this outcome.

The reduction in lenient responding produced by the evaluative format items was not large enough to result in a significant increase in rater reliability. No significant differences were found in rater reliabilities among the three response cue formats.

In all, it was found that the study was partially successful in obtaining its ends--succeeding in reducing lenient responding by changing the response mode, but failing to reduce it sufficiently to improve the rater reliability of the instructional rating form items. Since the evaluative format items of the instructional rating scale were found to be least prone to leniency bias, comparable in rater reliabilities to Likert and descriptive

formats, and most consistent with the experiences of the raters and the normative purposes of the rating task, it was concluded that they were the best choice of the three formats to improve the existing instructional rating scale.

ALTERNATIVE RESPONSE DEFINITIONS IN INSTRUCTIONAL RATING SCALES

By

Barbara Houghton Showers

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Personnel
Services, and Educational Psychology

1973

DEDICATION

To

Donald and Lucille Houghton

ACKNOWLEDGMENTS

There are many people who have contributed directly and indirectly to this thesis. I would especially like to thank Dr. Robert Ebel, Chairman of my Guidance Committee, for his advice and counsel throughout my doctoral program; Dr. Leroy Olson whose ideas and professional experience were contributed at many points in the development of the thesis; Drs. William Mehrens, Robert Craig, and Stephen Yelon for their contributions as members of the committee; the instructors in the departments of Education, Humanities, Natural Science, and Social Science who volunteered to take part in the study; and all the members of the Office of Evaluation Services for their cordially contributed aid in developing the forms, scoring them, and making the data ready for analysis.

The financial support of the U. S. Office of Education through a Research Directors Training Program fellowship enabled me to complete my doctoral studies at Michigan State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

THE PROBLEM

Introduction

Over the years many efforts have been made to improve student ratings of teacher effectiveness. This study represents another such effort. It is concerned with the particular problem of the leniency bias shown by many students in rating their instructors. By leniency bias is meant the tendency of students to use only the two or three highest options in rating their instructors. The harmful effect of this bias is to reduce discrimination between instructors to the extent that small differences in mean ratings produce large differences in reported rankings. The idea which gave rise to the present study was that leniency bias could be reduced by changing the wording of the response options. It was hoped that a different wording would increase the range of options used by student raters and improve discrimination between instructors. The setting of the study is described below, first from the broad perspective of purposes, rationales, and problems of student evaluation of instructors, and second from the more specific perspective of events leading to this study.

Purposes, Rationales, and Problems

The purposes that student ratings can serve are threefold. They can be called normative, diagnostic, and informative (Gillmore, 1972). The normative purpose is served when the results of student evaluations are used by the department to help decide promotions, salary increases, and teaching assignments. The diagnostic purpose is served when an instructor makes use of the results of the student evaluation to improve his course. The informative purpose is served when the results of student ratings are made available to other students as they make decisions about selection of courses and instructors.

While student evaluation of instruction can be carried out in many ways, it is frequently accomplished by means of a single all-purpose questionnaire or rating form. It is recognized that the questionnaire is not always the most direct or most informative source of information to all instructors in all disciplines, but it is probably the most often used.

A review of some student instructional rating systems currently in use in the universities shows several variations on the themes described above. Some universities stress primarily the diagnostic purposes for instructor self-improvement, while others see student ratings as inputs into a larger system for departmental accountability. A review of seven rating forms indicates that all are used

for instructor self-improvement, three are used for some form of departmental accountability, and two may be used to aid students in selecting courses. A closer look at their rationales is presented below.

Instructor Improvement (Diagnostic Uses)

The Office of Evaluation Services of the University of South Florida concludes that student ratings of instruction are appropriate for instructor self-improvement but not for helping determine salary advancement or tenure. Its conclusion was based on the finding that there was a variation in average ratings between courses and departments, preventing one overall scale from being applied to all faculty. It felt ratings to be valuable for self-improvement, however, since, "If the instrument is designed to measure opinion of teaching functions and reliability is established, then its validity is assumed" (Caldwell, 1971, p. 3).

An even more cautious stance is taken at Southern Illinois University-Carbondale. Thomas Tyler of the Testing Center there suggests that in order to build a tradition for evaluation, instructor rating should be presented in a very non-threatening manner. The results should go only to the instructor who may then, if he wishes, release them to the department chairman or the student publication. Consistent with this philosophy,

the SIU form includes some forced choice items of the "non-evaluative" type, such as:

The one thing this instructor did best was to:
a. deliver good lectures.
b. encourage class participation.
c. understand and sympathize with students.
d. prepare a well organized course.
e. make good quizzes and examinations.

This type of item provides information to the instructor without a good-bad connotation (Tyler, 1972).

At Northwestern University, student evaluation of instruction is carried on by an outside agency, namely Educational Testing Service, which was asked by the Associated Student Government of Northwestern in 1970 to develop a questionnaire to gather student ratings of courses and instruction. The resulting instrument, SIR (Student Instructional Report), is now being marketed commercially by ETS (Centra, 1972). The primary goals of this instrument are teacher self-improvement feedback and provision of a high quality source of information for published student critiques of courses and instruction.

The use of the Purdue Rating Scale for Instruction is described in the manual as primarily for instructor self-improvement feedback. The writers stress the voluntary and confidential nature of the use of ratings, but note that 65% of the instructors in a study felt themselves benefited by the ratings and that 83% of the total sample among students, instructors, and administrators expressed belief that additional improvement would be possible with

continued use of the scale. The manual of the scale does provide comparative data in the form of percentile ranks, but use of the scale for departmental evaluation is not encouraged.

Departmental Accountability
(Normative Uses)

Unlike most of the scales developed for diagnostic uses only, the University of Illinois scale was developed with the philosophy that measurement is more useful when comparative results are available. When an instructor administers the scale, his results are compared with other instructors of his own academic rank, with those at the same course level, with other instructors in his particular department or college, and with all courses at the university. A shortened form of the original is being made available containing general summarized questions specifically designed to be used by departmental decision-makers to evaluate instruction. It is hoped that this form will become one input into a total instructional evaluation scheme (Aleamoni, 1972).

Student ratings of instruction are one of three inputs into the Faculty Appraisal System at Bowling Green University. They are the primary "point of view" by which the teaching dimension of faculty activity is evaluated. The system attempts to get the people closest to the activity to be the raters instead of placing full

responsibility in the hands of the department chairman; thus, students are the primary raters of teaching, while faculty peers and the chairman rate scholarly productivity and service (Swanson and Sisson, 1971). This system uses the University of Illinois scale to gather student ratings since the scale objectives are compatible with those of the system.

The use of student ratings of instruction has been made a mandatory procedure by the Academic Council at Michigan State University. In 1969 the Council approved the following procedures "as a means to assist in improving the evaluation of instruction. . . .

a. Each of the teaching faculty (including graduate assistants) at MSU regardless of rank or tenure is required to use the Student Instruction Rating Report to evaluate at least one course in every quarter in which he teaches and every separate course he teaches at least once a year.

b. The results generated by the Instructional Rating Report shall be evaluated at the departmental level in order to help determine individual effectiveness. Appropriate procedures for the execution of this evaluation shall be determined according to departmental or residential faculty prerogatives.

c. The department chairman will be asked to describe in his annual report the steps which have been taken by the department or residential college to improve instruction (MSU Faculty Handbook, 1971, p. 42).

The Student Instruction Rating Report is a machine-scored 21-item questionnaire on which normative data have been

developed. The instructor receives a printout of his rating results giving mean, standard deviation, and percentile ranks for each item and each subscale.

Informative Uses

A separate discussion of the use of ratings for student publications (Informative Uses) was not conducted since few of the seven universities utilized this function of student ratings of instruction. Table 1.1 presents a summary of the uses of instructional rating scales at the seven universities discussed.

TABLE 1.1.--Uses of instructional rating forms in seven universities.

University	Normative[a]	Diagnostic[b]	Informative[c]
Bowling Green	X	X	
Michigan State	X	X	
Northwestern		X	X
Purdue		X	
SIU-Carbondale		X	optional
Illinois	X	X	
South Florida		X	

[a]Comparisons are made with other instructors.

[b]Instructor uses results to improve instruction.

[c]Results are made available to students to choose courses.

Status and Deficiencies of Student Ratings

There are many ways to evaluate instruction other than by the use of student ratings. These are, for example, the evaluation by chairmen of departments, by deans, by colleagues, by alumni, by amount and quality of scholarly research and publication, by informal student opinion, by committee evaluation, grade distribution in classes, student examination performance, enrollment in elective courses, course syllabi and exams, classroom visits, and other more informal methods. Why do many universities use student ratings as at least one input into their teacher evaluation systems? Spencer and Aleamoni at the University of Illinois suggest that since the students are the prime beneficiaries of the instruction, they appear to be the most logical evaluators of the quality and effectiveness of the course elements:

> In addition, student opinions should indicate areas of rapport, degrees of communication, or the existence of problems and thereby help instructors as well as educational researchers describe and define the learning environment more concretely and objectively than they could through other types of measurements (1969, p. 1).

Remmers and Weisbrodt of Purdue take a somewhat dimmer view of students' capabilities but advocate student ratings for this reason:

> Whether the student's judgment is correct is largely beside the point.The real point is that his attitude toward the instructor is a vital factor in the total learning situation. . . . Nor has the teacher any choice as to whether he

> will be 'rated' by his students. Such rating goes on in every classroom everywhere. The only real choice the instructor has is whether he wants to know what these ratings are. If he chooses to get this knowledge, he is in a position to profit thereby. He will have obtained the possibility of control of one of the important elements in the total learning situation (1965, p. 1).

These two statements illustrate the primary arguments for the use of student opinion in the evaluation of teachers. Students are the day-to-day consumers of the instructional product and in addition, the success of instruction appears to depend on their positive attitude toward the learning environment.

Validity of student opinion.--Although students have the most opportunity to observe instruction, questions have been raised about the influence of such variables as the student's sex, GPA, major, or personality on his or her ratings of instruction. The majority of the review of student rating research conducted by Costin, Greenough, and Menges in the Spring, 1971 Review of Educational Research is devoted to this topic. It appears that few strong or consistent relationships have been found between student demographic variables and student opinion of instruction, indicating that student opinion is not apt to be biased by factors other than the instruction received. After reviewing some fifty studies on the subject, Costin et al. (pp. 520 and 530), concluded:

1. "Correlations between course rating and grade received, when observed at all, tended to be small."

2. "Majors tended to rate courses more highly than non-majors in some cases."

3. "Students required to take a course sometimes rated it lower than those for whom it was an elective."

4. "Upperclass students occasionally gave higher ratings than underclassmen."

5. Experienced or higher ranking instructors usually received higher ratings than did their less experienced colleagues."

6. "A number of studies found no significant differences in overall ratings of teaching made by men and women students, or in the ratings received by men and women teachers."

John Centra noted in the Student Instructional Report (ETS) three additional points:

7. "Students on campus and alumni agree on average ratings of the same instructors" (r's between .40 and .68).

8. "Student needs (as measured by the Edwards Personal Preference Schedule) were found to influence some items on the Purdue Rating Scale (Rezler, 1965)."

9. "When teacher personality measures and student ratings have been correlated the relationship has been generally negligible (Borg, 1957; Bendig, 1955)." However,

both Centra and Costin et al., suggest this area has not been conclusively researched.

University of South Florida researchers, Remmers of Purdue University, and Wilson and Hildebrandt in California found in their research an additional relationship:

10. Differences in ratings can be expected between departments or courses within specific colleges.

Few demographic variables had consistent, strong effects on student opinion in the research reviewed. Only the experience of the instructor and department affiliation repeatedly appeared to show differences in ratings.

While student demographic characteristics have not been shown to bias their ratings of instruction in most cases, further questions have been raised regarding the validity of the rating form itself as a criterion of teaching effectiveness. Concerns have been expressed over (1) deficiency of the rating form alone as a criterion of teaching effectiveness, (2) contamination of ratings by halo effect and question ambiguity, (3) scale unit bias due to "generosity errors," and (4) criterion distortion by improper weighting of results. Investigations of these concerns suggest that the validity of the rating form as a criterion of teaching effectiveness depends on its proper use with other inputs to teacher evaluation and on its susceptibility to scale unit biases. The framework for

the critique of the rating form as a criterion measure is provided by Brogden and Taylor (1950), who originally outlined the four types of bias described above as possible criticisms of any criterion measure. Studies pertaining to each criterion bias as it relates to student ratings of instruction are detailed below. Later in the report, possible characteristics of a good instructional rating scale are derived from this discussion.

Criterion deficiency.--Several authors make the point that student ratings of instruction would be deficient as the sole criterion of teaching effectiveness, but it has been demonstrated also that student ratings represent one stable part of such a criterion. Costin et al. (1971), report that students repeatedly cite: (1) knowledge of subject, (2) organization of course content, (3) enthusiastic attitude toward teaching and subject, and (4) interest in students, as attributes of most importance in teaching effectiveness, but the correlations between student ratings and faculty peer ratings or department chairmen's ratings in various studies ranged from .08 to .63. It appears that student ratings are a stable but relatively independent part of a larger criterion of teaching effectiveness.

Criterion contamination.--Major measurement texts often cite halo effect and ambiguity of the quality to be observed as major influences affecting the rater's

ability to rate accurately in any situation. Halo effect was not often mentioned in studies of student ratings of instruction, but where it was investigated (Remmers, 1934; Hodgson, 1958), little influence was found. Several authors offered general suggestions for wording a rating scale in order to reduce ambiguity of the quality to be observed, but none specifically considered student ratings of instruction. Cronbach (1960) suggested that such words as "average" and "excellent" be replaced by specific descriptions of behavior. Both he and Thorndike and Hagen (1961) suggested that abstract terms such as "leadership" or "personality" not be rated, but rather more overt, directly observable characteristics be rated, such as "pleasant speaking voice," or "appearing at ease at social gatherings." Oppenheim (1966) spoke of defining a frame of reference in a rating scale with much the same intent--to make sure every judge agrees on the meaning of the trait to be rated. He claimed the increased specificity of traits to be rated also would tend to decrease the halo effect by making raters less able to generalize their ratings.

Criterion scale unit bias.--Scale unit bias seems to be a particular problem of all rating scales. Piling up of ratings at the upper end of the scale, failure to employ lower scale units, piling up in the center of the scale have all been frequently reported in research with

rating scales. The "generosity error" causing piling up of ratings at the upper end of the scale is a persistent problem when people rate other people. As Thorndike and Hagen put it, "There seems to be a widespread unwillingness on the part of raters to damn a fellow man with a low rating" (1961, p. 344). This is apparently true of students in their evaluation of instructors. Two examples of generosity error in an instructional rating scale can be taken from forms administered at Michigan State University and the University of Iowa. At MSU the average rating on SIRS in 1971 fell between 1.7 and 2.5 on a five point scale (Office of Evaluation Services, MSU, 1969), and at Iowa the average rating on a 1952 experimental form ranged from 1.5 to 2.6 on a five point scale (Stuit and Ebel, 1952). Attempts have been made to counteract generosity error by manipulating response options. Evaluators at SIU-Carbondale report some success using a favorable midpoint on a five point scale and using no disparaging options (instead of "terrible" use "needs considerable improvement") in order to encourage raters to use the full range of the scale (Tyler, 1972). Amiel Sharon (1970) developed a forced choice student instructional rating scale and discovered that choosing between two to four equally favorable statements to describe the instructor was resistent to bias but it could no longer produce a profile of the instructor's strengths and weaknesses.

Criterion distortion.--Criterion distortion arises out of the improper assignment of weights to the several elements of a criterion measure. The relative importance of student ratings of instruction among all inputs in evaluating teaching effectiveness depends on the particular philosophy of the college or department. As Brogden and Taylor point out, adequate empirical procedures for deriving any criterion combination have yet to be developed.

In summary, the validity of student ratings as a criterion of teaching effectiveness depends on their use in combination with other inputs, such as faculty opinion and department chairman's evaluation, and on their susceptibility to halo effects and generosity errors. Research cited showed that halo effects were apparently minimal with student rating of instruction and that generosity of ratings might sometimes be discouraged by manipulating the wording of the response options.

Reliability of ratings.--To complete the background information available to one considering a study of student ratings of instruction, studies to date concerning the reliability of student ratings will be summarized. Most have reported moderate to high coefficients, indicating that reliability is not a serious problem in this area of rating scale construction.

Several approaches were taken to measuring the reliability of student ratings of instruction. Some

considered the stability of ratings over time, others obtained internal consistency estimates via Cronbach's alpha, and still others considered item and rater reliability.

Early studies of stability of student opinion over time yielded correlations of .87 to .89 over periods of two weeks to a year (Costin et al., 1971). Recent studies confirmed the stability of student ratings over time (Costin, 1968; Wilson, Hildebrandt, and Dienst, 1971). Reported split-half reliabilities were .79 and .92 for two instructional rating scales (Lovell and Haner, 1955; Spencer and Aleamoni, 1969), while use of Cronbach's alpha on subscales produced internal consistency reliabilities ranging from .58 to .985 in studies of five instructor rating scales (Gillmore, U. of Illinois, 1972; Hildebrandt, Wilson and Dienst, 1971; Centra, Educational Testing Service, 1972; MSU Technical Report, 1969; Tyler, SIU-Carbondale, 1972). Item reliabilities reported for four scales ranged from .40 to .96 with median values greater than .80 (Gillmore, 1972; Remmers and Weisbrodt, 1965; Coffman, 1954; and Deshpande et al., 1970).

It can be seen that student raters report their opinions with moderate to high consistency within the form and with high stability over time. It appears that their opinion can be trusted to be more than a whim of the moment

and thus could prove useful to an instructional evaluation system.

Psychometric Characteristics of an Ideal Student Instructional Rating Form

The functions, rationales, and problems of student rating of instruction have been described, and the progress of reliability and validity studies has been discussed. It now remains to discuss the psychometric qualities of an ideal instructor rating scale and to describe the problems with an existing scale that led to undertaking this study.

The psychometric qualities of an ideal scale can be determined by keeping in mind the purposes for which the scale is used and the possible pitfalls to scale construction that were noted in the criterion validity studies.

From the functional point of view, when the scale results are used normatively by the department to help decide promotions, salary increases, etc., the ratings for good and poor instructors should be as widely different as possible to clearly distinguish between the recipients and non-recipients of the commendations. Psychometrically, the scale should discriminate between good and poor instructors for normative uses.

When the scale results are used diagnostically by the instructor to discover areas of teaching difficulty, the individual item mean ratings should represent close

agreement among the students in the class on each trait that each item represents. In most rating forms each item concerns one aspect of teaching, such as course organization or student's opportunity to ask questions. It should be possible that students agree on a high rating for one item and also agree on a low rating for another item, thus giving a clear direction to the instructor for self-improvement. The combined psychometric attributes of discrimination between good and poor instructors and close agreement of students within a given class on each item can be measured by the intraclass rater reliability coefficient. This statistic compares the amount of variation in ratings between instructors for an item with the amount of variation in ratings within each instructor's class for that item. If there is as much variation in the ratings given to a single instructor by his students as there is between the scores of all instructors rated, then the statistic returns a value close to zero, indicating that the differences between good and poor instructors are indistinguishable from the difference in the student opinions of one single instructor. This could happen when the item is so ambiguous that the students are unable to agree on its meaning, or when both good and poor instructors are so alike on a particular trait that it is not useful to include it for diagnostic or normative uses. A high rater reliability for an item would indicate that

it was both discriminating and unambiguous--at least to the student raters whose opinions were being sought.

Whether the scale is used for normative or diagnostic purposes, it must be kept in mind that the task of the scale is to solicit judgments from untrained student raters. Each question must contain enough information to make its intent clear but not so much that the rater is unable to digest it on the first reading. The format of the questions and response options becomes important in helping the rater to digest the information given and in helping him to return a response that reflects his true opinion. When the rating scale format is such that the rater finds himself always making the same response, even to widely different questions, then the format is encouraging response set biases. Three major types of response bias possible with rating scales are leniency (same as "generosity error"), central tendency, and halo effect. Leniency bias occurs when raters use only the high response options on a scale, central tendency occurs when raters use only the middle options, and halo effect occurs when a rater rates all traits of one ratee alike because of his general impression of the ratee. Instructor rating scales have been found to be most susceptible to the leniency bias of high ratings of all instructors. Such a bias works against ability to discriminate between good and poor instructors as well as perhaps against the validity of the rating form

itself. The amount of leniency bias in a given response cue format can be measured by finding the closeness of the mean of all instructors rated to the midpoint of the response scale, assuming that teaching ability is normally distributed about the midpoint of the scale. Inspection of the variation in ratings about the midpoint could rule out the presence of central tendency in this situation.

Impetus for the Study

The ideal student instructional rating scale would be free of response biases, discriminate between good and poor instructors, and have unambiguous questions on which all raters could agree for each instructor. Such a scale would have to be carefully developed from items selected for appropriate content as well as for their psychometric characteristics.

The impetus for undertaking this study was created when such a carefully developed scale was found to still possess a strong tendency to leniency bias in the ratings produced. Even after a substantial data base had been established over a five year period, the mean item responses on the scale ranged from 1.7 to 2.5 on a five point continuum where 1 is the highest rating. This essentially psychometric problem was compounded by instructor confusion in interpretation generated when (1) the results were reported to instructors in percentile ranks, and (2) a University policy was approved whereby every instructor was

required to use the rating form in at least one course he taught every term and report the results to his department chairman. The confusion became apparent when the leniency of the ratings on the scale caused reports of performance at the 50th percentile or below to be given to instructors whose classes only "agreed" and did not "strongly agree" to some of the statements in the questionnaire. This occurred because the distribution on which the percentile ranks were based (the five year data base) was centered about the high item mean rather than about the midpoint of the five point continuum as one would conventionally interpret an "average" value. Thus, when the 50th percentile rank was legitimately assigned to the mean value of the item and lower ranks were given accordingly, an instructor with a score less than the mean would be given a less than 50th percentile rank even if the score was still well above the midpoint of the five point scale (Figure 1.1). It appeared that a study was needed to attempt to reduce the leniency of students' responses so that the 50th percentile could be more conventionally interpreted as a midscale value.

For purposes of such a study, it was not considered productive to create a new set of items for the scale, since discovering the appropriate content was well done in 1967 when an elaborate selection system was set up to determine what questions were to be on the form. Faculty

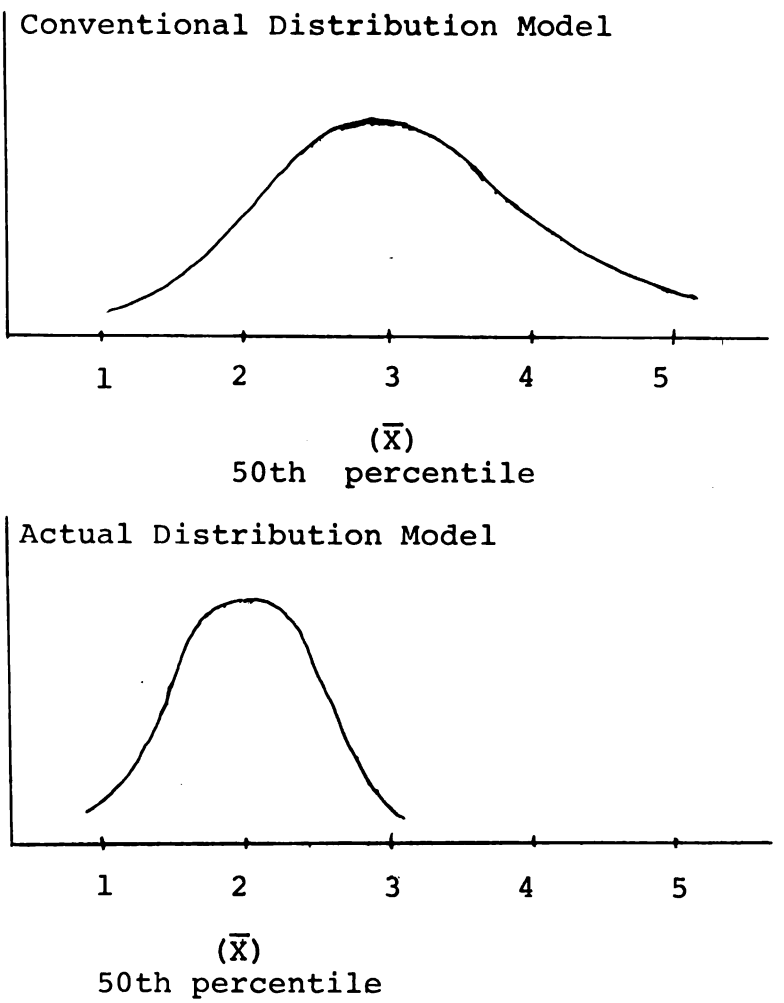

Figure 1.1.--Illustration of the differences between the actual norm group distribution and a conventional norm group distribution.

and students were polled as to the usefulness and appropriateness of a large pool of potential items, and 56 items with the greatest concensus of favorable opinion were pretested, yielding the 21 most discriminating in five areas that became the final scale. The recent Review of Educational Research (1971) review of studies of teacher effectiveness showed that most instructor rating scales developed had at least four of the five factors present in this scale, indicating that there is agreement at what constitutes teacher effectiveness.

Given the well-established content of the scale, it seemed more reasonable to look at the possible effect of manipulation of response options in reducing the leniency bias problem than to revise the entire scale. It was hypothesized that since altering response cues had reportedly reduced leniency response bias in some studies (Smith, 1967; Stockford and Bissell, 1949; Guilford, 1954; Cronbach, 1950), it might do so here, and might also improve rater reliability of the student instructional rating form by increasing the amount of scale used so that there could be maximum latitude for discrimination between instructors. Thus, a study was devised in order to compare the abilities of different response cues to reduce leniency of response and to improve rater reliabilities of the items. The choice of response cues was to be based on attributes of the different response cue types reported

in the literature. Rater reliabilities were calculated because the focus of this coefficient was on consistency of student agreement in ratings and on the students' ability to discriminate between instructors when using a particular response cue format.

Purpose

The purpose of the study was to compare the effects of alternate response definitions on the leniency and rater reliabilities of student instructional rating form items. After a review of the literature on response cue types, three response formats were selected for the study. They were defined as, (1) fixed alternative Likert cues (SA-SD), (2) fixed alternative evaluative cues (superior-inferior), and (3) multiple choice short descriptive cues. In addition to finding the least biased and most reliable of these item types for student rating of instruction, the study tested two claims concerning leniency bias that were made in the literature:

a. Evaluative cues are more susceptible to response bias than other cues.

b. Fixed response alternatives are more susceptible to bias than descriptive multiple choice alternatives.

Hypotheses

Null Hypotheses

H_1. There are no differences in mean ratings of instructors between items with Likert, evaluative, and descriptive response cue formats.

H_2. There are no differences in rater reliabilities between items with Likert, evaluative, and descriptive response cue formats.

Alternative Hypotheses

H_{1a}. The mean ratings with the evaluative format will be significantly more lenient than the mean ratings for the Likert and descriptive formats.

H_{1b}. The mean ratings with descriptive cue formats will be significantly less lenient than the mean ratings for Likert and evaluative cue formats.

H_{2a}. The descriptive response cue format will have significantly more rater reliability than the Likert or evaluative cue formats.

Summary of Response Cue Literature

The literature on response cues is presented in detail in Chapter II, but it is briefly summarized here to help explain the selection of response types for the study.

The major part of the literature on response cue types does not deal specifically with student ratings of instruction. Rating scales have been more often used for sociological studies of behavior, personnel evaluation, and psychological or vocational counseling. But some generalities appear to have emerged from these diverse uses that might be expected to hold in the student rating situation.

To begin with, there are several generally accepted formats for response cues. Each type provides the rater with a slightly different task, though the different types have been used interchangeably in the same rating situations. As Guilford (1954) defines them, they are: numeric, descriptive graphic, standard, cumulated points, and forced choice. The numeric scale provides a number continuum from which the rater assigns a number value to a ratee's trait or behavior, while the descriptive graphic scale adds descriptive words, sentences, or paragraphs to define points on the continuum, and the rater chooses the description that best fits the ratee. A standard (evaluative) scale provides a real or assumed norm group against which to compare the ratee and the rater's task is to judge whether the ratee is average, above average, etc., with respect to the group. The cumulated points (Likert) scale provides several statements to which the rater agrees or disagrees in varying intensities. His

responses are then summed to arrive at his overall opinion of the ratee. The forced choice scale provides the rater with two-to-four equally favorable or unfavorable statements from which the rater must choose the ones most descriptive of the ratee.

Many authors hypothesize that some of these tasks, when done repetitively in a questionnaire, are more susceptible to response biases than others. For example, Cronback (1950) distinguishes between these scales according to whether the alternatives are the same or different for every question rated. He opines that raters are less likely to develop a set response when faced with different alternatives every time than when faced with the same alternatives for every question. (Fixed alternative response options would be presented by numeric, standard, and cumulated points scales, while multiple choice options would be presented by descriptive graphic and forced choice scales.) Other authors cite other characteristics of the response options which they hypothesize could make the rating task more or less susceptible to response bias.

Most discussions of rating scale techniques dwell on practices in avoiding response set with the various response option types. These practices include manipulating extremeness of cues, direction of scales, spacing of cues along a continuum, balance of favorable and

unfavorable cues, presence or absence of neutral or undecided cues, and concreteness of descriptions of cues. Evidence reported in the literature on these practices is summarized below. Where the evidence for some statements is contradictory in part, or non-experimental, the statements are listed as claims to be further tested. Details concerning the studies contributing evidence to each statement are presented in Chapter II.

Summary Statements

1. The optimal number of options for each question is five to seven when untrained raters are used.

2. The presence of a neutral point increases the ambiguity of the scale.

3. Reduction in leniency bias due to reversing the direction of the scale within a questionnaire may increase the errors in rating.

4. Leniency bias may be reduced by the presence of more favorable than unfavorable response options.

5. Numeric, sentence, or paragraph cue lengths may reduce leniency bias, if the cues are not too long, but cue length has no apparent effect on the rater reliability of untrained groups of raters.

Claims to be Further Tested

1. Evaluative cues are more susceptible to response bias than other cues.

2. Fixed response alternatives are more susceptible to bias than descriptive multiple choice alternatives.

3. Reported rater reliabilities for instructor rating scales currently in use roughly rank the cue types in decreasing order as descriptive (.87 and .86), Likert (.84), and evaluative (.81).

The study tested the two claims concerning leniency bias in a controlled setting by comparing the means of the Likert, evaluative, and descriptive response cue formats in equivalent rating situations with student evaluation of instructors. The choice of cue types was guided by the derived results above. In order to study the question of bias in a controlled setting, the number of options for each cue type was held constant at five, within the range of optimally reliable numbers of options for untrained raters. Although the scale to be improved was a Likert scale having a neutral midpoint, the alternative scales did not have Neutral as an option in order to decrease the likelihood of ambiguity. None of the questions compared were stated in the opposite direction from the others. (Four such questions existed in the scale to be improved but were omitted from the analysis for this and further reasons--see Chapter III.) The balance of favorable and unfavorable cues was held constant in this study in order to isolate the effect of cue type (Likert,

evaluative, descriptive) on lenient responding. Likert cues were chosen for the study because the scale to be improved was in Likert format. Descriptive cues were chosen because they were the most often recommended to reduce rater biases (in spite of the few negative findings reported). Evaluative cues were included to test the claim that they were the most bias-prone cue type. Rater reliabilities were also compared across the three response formats where previously only comparisons with numeric formats had been made experimentally, and the descriptive format was hypothesized to have the greatest rater reliability. This hypothesis was based on the assumption that the descriptive format would be the least bias-prone and would produce greater rater reliabilities than the other formats by improving discrimination between instructors. The rater reliabilities of existing scales seemed to concur with this prediction.

Overview

In Chapter II, the literature on response cues and rater reliability is reviewed in detail. The design and procedures of the study are discussed in Chapter III, and the results concerning the leniency bias and rater reliability of the three response cue formats are presented in Chapter IV. Conclusions and discussion of the results appear in Chapter V along with a summary of the study problem, theory, and methodology.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

The studies of response cues reviewed in this chapter provided the theoretical groundwork for the selection of the three response cue types compared in this study. The studies of rater reliability also reviewed in this chapter provided the information necessary to formulate and test the hypothesis concerning rater reliability of the instructional rating forms. Studies concerning the background and setting of the problem itself were presented in Chapter I.

The major part of the literature on response cues and on rater reliability does not deal specifically with student ratings of instruction. Rating scales have been more often used for sociological studies of behavior, personnel evaluation, and psychological or vocational counseling. But some generalities appear to have emerged from these diverse uses that might be expected to hold in the student rating situation. The studies are presented in detail in the following sections under the headings, "Studies of Response Cues," "Data on the Reliability of Cue Types," and "Studies of Intraclass Rater Reliability."

The Summary section which follows presents the generalities derived from the studies and summarizes the evidence contributing to each statement.

Studies of Response Cues

Types Detailed

Guilford (1954) defines five broad categories of response cues: numeric, descriptive graphic, standard, cumulated points, and forced choice. Similar categories are defined by Thorndike and Hagen (1961) under corresponding titles: frequency of occurrence or typicality, behavioral statement, man-to-man, and present-absent. They add percentage of group and ranking to the list. Oppenheim (1966) mentions Thurstone and Likert type scales whose response options would probably fit into the "present-absent" and "cumulated points" categories already mentioned. Levinthal et al. (1971), discuss a scale format of real-ideal discrepancies, and Cronbach (1950) distinguishes between multiple choice and fixed format response cues.

On Response Set

Most discussions of rating scale techniques dwell on practices in avoiding response set with the various response option types. These practices include manipulating extremeness of cues, direction of scales, spacing of cues along a continuum, balance of favorable and

unfavorable cues, presence or absence of neutral or undecided cues, and concreteness of descriptions of cues. The diverse uses of rating scales led different researchers to study these practices in different contexts, however; hence, the rating situations and the variables manipulated by the experimentor are inconsistent from one study to the next--from foremen rating subordinates to mental patients rating their self-concepts. But they indicate what manipulations of response cues have been made, and their outcomes. The studies are categorized here according to the practices on which they provide data. Their results are further condensed into a series of general statements appearing in the Summary section of this chapter.

Number of options.--The effect of number of options on leniency can be seen in a study by Hillmer reported by Edwards (1970). After administering a nine-point scale, Hillmer selected the two options on either side of the item median and readministered the two-choice scale. Instead of an equal distribution of choices about the median, 73% chose the higher of the two options given.

Direction of scale.--Elliott (1961) tested Likert items on the same positively and negatively worded topics and found that tendency to agree with the direction of the statement was apparent for middle and low aptitude subjects, but not for high aptitude subjects, whose scores remained relatively stable.

Madden and Bourdin (1964) compared orientation and numbering of nine-point scales and found statistically significant differences between the scale means. The greatest difference seemed to be between the horizontal graphic scale numbered 1 to 9 which produced the least lenient ratings, and the vertical scale numbered +4 to -4 which produced the most lenient ratings. But no means fell below the scale midpoint.

Reversing directions of scales within a questionnaire is argued on intuitive grounds by Oppenheim (1966) that it forces raters to stay alert and doesn't allow them to create a habit of marking every question in the same place. On equally intuitive grounds, Guilford (1954) claims that reversing scale directions generates more rater errors than it does unbiased responses.

Spacing and balance of cues.--Regarding spacing and balance of favorable and unfavorable cues in a graphic scale, Guilford notes, "To counteract leniency error, the cues on the favorable side may be more widely spaced and more numerous than those on the unfavorable side" (1954, p. 268). In practice, Tyler (1972) chose a favorable midscale anchor making three favorable cues out of five in the SIU instructor rating form and still found that few mean ratings fell below item midpoints, though some reductions were obtained.

Follman (1973) carries Guilford's advice to the extreme in comparing the conventional five-point balanced evaluative scale having two favorable options to three other five-point scales each having one more favorable option than the last. The most favorable scale created was, "Above Average"; "Superior"; "Excellent"; "Superb", "Perfect." The students gave the instructor a mean rating between the first and second highest options ("Above Average" and "Superior") on the conventional scale, and between the second and third highest options on each of the succeeding more favorably weighted scales. It appeared that a favorable midpoint helped reduce leniency bias, but that more favorably weighted scales had little further effect on leniency.

Other approaches to cue balance and spacing tend to favor equally weighted cue distributions. Champney (1941) favored equal spacing and balance of cues to the extent that he devised a pretest of cue placement akin to the Thurstone equal-appearing intervals technique that allowed him to determine a scale value for each cue on the continuum and pick out unambiguous, equally spaced high, medium, and low cues for the final scale.

Amiel Sharon (1970) found he was able to avoid leniency bias in student ratings of instruction by using forced choice scale items which balanced favorable statements against each other, but he notes that it could

not be used for diagnostic purposes since it only gave a single overall score for each instructor.

Presence or absence of neutral.--Regarding the presence or absence of a neutral point on the cue continuum, Guilford and Jorgensen (1938) found a tendency to bimodality in distributions which they thought were unimodal. This was more serious with the numeric than the graphic scale. Since the point of lowest frequency in the numeric scale was at the indifference category, they suggested eliminating the indifference category in numeric scales and not mentioning indifference in a graphic scale except as attached to a point.

Cronbach, in "Response Sets and Test Validity" (1946), opts for those practices which will reduce ambiguity, one of which is, in his judgment, eliminating the neutral response option.

Holdaway (1971) found results contrary to those of Guilford and Jorgensen in his study of response distributions in a Likert scale with and without a neutral point. His distributions peaked at the "Agree" option and declined on either side whether a neutral point was present or not. But a greater percentage chose the disagree option when no "Neutral" choice was available, or when the N was placed after the SA-SD scale.

Concreteness of cue descriptions.--Both Cronbach (1946) and Guilford (1954) stress the importance of clarity and specificity of cues. Guilford states, "Avoid using cues of a very general character, such as 'excellent,' 'superior,' 'average,' 'poor,' and the like" (1954, p. 293). But Symonds (1931) points out that the difficulty of vocabulary should be considered, taking care to avoid unusual words even though they are highly descriptive and meaningful, such as "slovenly" for "very careless in dress."

Concerning lack of specificity of evaluational cues, Stockford and Bissell (1949) recount a study in which values from 1 to 100 were assigned by 200 raters to cues which could be used in a rating scale. The ranges and standard deviations of the values for those cues which contained evaluative words ("average," "excellent," etc.) were significantly greater than the ranges and standard deviations of the non-evaluative cues.

At one time it was thought that the man-to-man scale would provide the concreteness of description necessary to avoid leniency response set in an evaluative type of cue. But in their development of a man-to-man instructor rating scale, Stuit and Ebel (1955) note that the norms they derived all lay in the upper half of the five-point scale with an overall mean of 2.04 for 267 classes. The instructors may have been a select group, but the ratings were very high for such a large number of classes.

The effect of multiple choice versus fixed alternative cues on leniency bias has been studied in several ways, with uncertain results. Smith reports that acquiescence response set is best dealt with by constructing items that avoid the agree-disagree format in favor of "contentful alternatives" (Smith, 1967, p. 88), but doesn't substantiate his claims. Similarly Cronbach hypothesizes that multiple choice items are least susceptible to bias. He states:

> Item forms using fixed response categories are particularly open to criticism. The attitude test pattern, A, a, U, d, D, is open to the following response sets: Acquiescence . . ., evasiveness . . ., and tendency to go to extremes. . . . (1950, p. 21)

Elliott (1961) claimed this was not the case in her study where most acquiescence occurred with items in multiple choice rather than fixed alternative format, but she did not make the items more descriptive than the existing Likert alternatives restated in sentence form.

Champney (1941), in his work with the Fels Parent Behavior Rating Scale, opts for long cue explanations if the raters are trained but short cue explanations if the raters are not. Bryan (1944) appeared to confirm this opinion with untrained student raters when he found no difference between mean ratings of given instructors when the cue alone was used (excellent, good, average, etc.) and when the cue followed by a paragraph explanation was used. Finn (1972), also using untrained raters, found no

differences in mean ratings between cues which were paragraph explanations and numeric cues. In the cases of both Finn and Bryan the paragraphs were several sentences long, rather than a few descriptive words. Stockford and Bissell (1949), on the other hand, found that errors of leniency were less for ratings made on sentence-length descriptive graphic scales than for those made on single word evaluative scales.

Data on the Reliability of Cue Types

Primary experimentation has involved increasing the number of response options to some optimally reliable point. Guilford (1954) discusses this research and concludes that five to seven options is a conservative choice, and that the optimal number to use depends on the ease of rating the trait and the training and motivation of the raters. Mattell and Jacoby (1971) point out that most research on this question has dealt with internal consistency measures. They found no differences in test-retest reliability of 2 to 19 option Likert scales using untrained student raters. But Finn (1972) confirms that five to seven options give optimal inter-judge agreement on each item with untrained student raters. (His formula for interjudge agreement: $r = 1 - \frac{\text{var (observed)}}{\text{var (random)}}$.)

Other experimentation has compared the rater reliabilities of various verbal cues to numeric cues. J. B. Taylor *et al.* claim from their previous research

that, "whereas numerical rating scales show a typical inter-judge reliability in the r = .40 to .60 range, example anchored scales typically show reliabilities in the .70 to .99 range--and this with untrained raters" (Taylor et al., 1972, p. 544). Their examples are short behavioral statements anchored to a point on a thermometer-like scale. Peters and McCormick (1966) found significant differences in single rater intraclass reliabilities between numeric and one sentence job-task anchored scales, but the differences vanished when the r's were stepped up by the Spearman-Brown formula to become the reliabilities of mean ratings from n raters. Similarly, Finn (1972) found no differences in stepped-up intraclass rater reliabilities between numeric and paragraph-length cues.

Some collegiate instructor rating scales report rater reliabilities. Since the scales use different cue types, it is possible to make a rough comparison of cue type reliabilities in this way.

Rater reliabilities are available for the Purdue scale (Remmers and Weisbrodt, 1965), Oklahoma A&M scale (Coffman, 1954), Georgia Tech scale (Deshpande et al., 1970), and U. of Illinois scale (Gillmore, 1972). The first is descriptive graphic in part and evaluative in part, the second is descriptive graphic, the third is a five-point frequency of occurrence scale, and the fourth is a Likert type with no neutral option. The median reliabilities are

.87 and .86 for the descriptive graphic scales, .84 for the Likert type scale, .81 for the evaluative scale, and .79 for the frequency scale. Numbers of raters averaged at least 20 per class in each calculation.

Studies of Intraclass Rater Reliability

Methods of estimating rater reliability are discussed by Ebel, Lindquist, Stanley, Cronbach, Rajaratnam and Gleser, Remmers, Medley and Mitzel, Guilford, and Brown, Mendenhall and Beaver. Most are analysis of variance procedures, predominantly the intraclass correlation coefficient. Medley and Mitzell (1963), Guilford (1954), and Brown, Mendenhall, and Beaver (1968) consider only the two-way analysis of variance case where instructors and raters are completely crossed in the design, i.e., where every rater rates all instructors. This design is not comparable to the student rating of instruction situation where it is unlikely that any rater rates more than one instructor in the study. Ebel (1951), Lindquist (1953), and Stanley (1971), allude to generalized intraclass reliabilities where the raters may be different for each instructor. Ebel concludes, after discussing three formulas applicable to rating situations--average intercorrelation (Peters and Van Voorhis), the intraclass formula, and the generalized formula for the reliability of averages (Horst)--that the intraclass correlation formula is most versatile, allowing one to include or exclude "between

raters" variance from the error term. (One would include between-raters variance in the error term in the student rating of instructors situation since all raters do not rate all instructors.) Also, as Engelhart (1959) points out, both a single rater estimate and an n-rater estimate can be obtained with the intraclass coefficient while Horst only gives the n-rater case. In addition, estimates of precision can be readily calculated from an intraclass correlation. Both Ebel and Lindquist explain how to calculate confidence intervals for the intraclass coefficient.

Cronbach, Rajaratnam, and Gleser (1963) explain how the use of the intraclass formula allows one to generalize from randomly selected samples of raters to the reliability of raters in general. This is particularly desirable in determining the reliability of student ratings of instruction, since the particular group of students who were rating each instructor is certain to be different every time.

The intraclass coefficient for the "average" rater can be stepped up by the Spearman-Brown formula to give the reliability of a number of raters (Stanley, 1971). Remmers provided empirical verification of this use of the Spearman-Brown formula in two often-quoted experiments with the reliability of student ratings of instruction. He concluded that judgments were equivalent to test items in the sense of the Spearman-Brown formula and that the

formula could predict within one standard deviation the reliabilities empirically obtained (Remmers, 1927 and 1931).

In another study, Remmers (1934) determined average rater reliabilities of a single rater for high school and college students for three items with 57 teachers. The non-stepped-up reliabilities reported for college students averaged .290 ± .102 for the "interest in subject" item, .429 ± .094 for the "presentation of subject matter" item, and .354 ± .038 for the "stimulating intellectual curiosity" item. These results seem to illustrate that the reported instructor rating form item reliabilities in the .80's and .90's are substantially affected by the number of raters assumed in the Spearman-Brown formula.

Summary

Most evidence reported here on the effects of cue types on response set and rater reliability can be categorized as either conclusions which most studies confirm or claims for which inconclusive or possible contradictory evidence was found. Summary statements are presented below, along with a review of the evidence contributing to each.

Summary Statements

1. The optimal number of options for each question is five to seven when untrained raters are used.

This conclusion is derived from the combined results of studies of rater reliability and studies of leniency bias. Guilford derives the five to seven estimate from his review of rater reliability studies and Finn specifically confirms with untrained student raters that five to seven options produce optimal rater reliability. While no such specific result is found with regard to the effect of number of options on leniency bias, Hillmer's example of the strong increase in leniency when the number of options was reduced from nine to two indicates the potential biasing effect of too few options on rater judgment.

2. The presence of a neutral point increases the ambiguity of the scale.

The studies of Guilford and Jorgensen and Holdaway support Cronbach's contention that the neutral response option causes ambiguity in rater responses. In Guilford and Jorgensen's study, raters avoided choosing the neutral option when it was the midpoint of a numeric continuum, while in Holdaway's study of the Likert format, the Undecided option was chosen if it was the scale midpoint but not if it was placed at the end of the scale. This variety of reactions to the neutral option supports the contention that raters are uncertain of its meaning in a rating scale.

3. Reduction in leniency bias due to reversing the direction of the scale within a questionnaire may increase the errors in reading.

Although Oppenheim argues on intuitive grounds that reversing question direction within a scale forces raters to stay alert, Elliott discovered that only the scores of high aptitude raters remain stable regardless of question direction while middle and low aptitude raters tend to agree with the direction of the statement. This supports Guilford's contention that most raters cannot be relied on to remain aware of the positive or negative wording of every question. The apparent reduction in leniency bias occurring by this method would seem to be largely due to the counteracting effects of acquiescence to both positive and negative statements.

4. Leniency bias may be reduced by the presence of more favorable than unfavorable response options.

Guilford's, Tyler's, and Follman's studies of the effect of the balance of favorable and unfavorable response options on leniency bias agree that the presence of more favorable than unfavorable options reduces lenient responding. However, Follman's results suggest that there is a limit to the amount of reduction in bias obtained by this method.

5. Numeric, sentence, or paragraph cue lengths may reduce leniency bias if the cues are not too long, but cue length has no apparent effect on the rater reliability of untrained groups of raters.

No differences in mean ratings were found by Bryan or Finn between short evaluative or numeric cues and long

paragraph explanation cues with untrained raters, but phrase-length descriptive cues were found in one study (Stockford and Bissell) to produce less lenient means than short evaluative cues, and were recommended by Champney for use with untrained raters. Studies by Peters and McCormick and Finn compared the rater reliabilities of numeric and sentence cues, and numeric and paragraph cues, respectively, and found no differences in rater reliability between these cue types.

Claims

1. Evaluative cues are more susceptible to response bias than other cues.

Evaluative cues are frequently used in measurement texts as an example of an ambiguous standard of reference that is especially susceptible to leniency bias. It is sometimes noted that in personnel ratings, an "average" rating is a condemnation of a person's performance. The texts would advocate less evaluative, more concrete cue descriptions to help counteract this tendency. But few studies have compared evaluative cues to other cues in the same rating task. Some supportive experimental evidence for this claim of greater susceptibility to bias among evaluative cues was found in Stockford and Bissell's paper in which they described two studies, one showing increased variance in ratings of the meaning of evaluative words compared with other non-evaluative words, and another study

showing more lenient ratings of subordinates by supervisors using evaluative cues compared to short descriptive cues. But possible contrary evidence was reported by Bryan when he compared evaluative cues with paragraph length cues in student rating of instruction and found no differences in mean ratings. It may have been the length of the cues which was at fault, but the finding disagrees with the generality originally made.

2. Fixed response alternatives are more susceptible to bias than descriptive multiple choice alternatives.

This is a broader claim than the one concerning evaluative cues, but since evaluative cues are a type of fixed response alternative, the studies applying to them also apply here. Thus, the Stockford and Bissell study comparing evaluative and descriptive cues confirms this claim, while the Bryan study casts doubt on it. When Cronbach published the claim, his primary objection to fixed response alternatives concerned the Likert cue type which he felt was prone to several biasing effects. Smith voiced a similar objection to Likert cue types in favor of descriptive multiple choice types ("contentful alternatives"). But another disconfirming result was reported by Elliott who compared Likert cues in fixed format to Likert cues in multiple choice format and found the most acquiescence in the multiple choice format.

3. Reported rater reliabilities for instructor rating scales currently in use roughly rank the cue types in decreasing order as descriptive (.87 and .86), Likert (.84), and evaluative (.81).

The values were taken from the technical reports of the various scales and were not obtained by experimental comparison of the three cue types, but this tentative rank ordering of rater reliabilities was compatible with the hypothesized order of bias-proneness of cue types. It seemed reasonable to assume that the least lenient cue type could allow greater discrimination between instructors than the more lenient cue types, and in this way produce the greater rater reliability. But this claim was yet to be tested.

One other topic on which studies were reviewed concerned the rationales and uses of the intraclass rater reliability coefficient. This topic was studied in order to aid in understanding and comparing the rater reliabilities that were calculated in the study. It was found that rater reliability could be predicted for any class size, and that confidence intervals could be generated about the single rater estimate so that comparisons could be made among them. It was further confirmed that the n-rater estimate made by means of the Spearman-Brown formula was a valid prediction of reliabilities empirically obtained.

In summary, it can be seen that there is no single statement that can encompass the results of all the studies

reviewed here. Rating scales have had so many diverse uses that there has been no unified line of research in the area. But, under close scrutiny, several generalities did emerge from the various studies which might prove useful to a study of student ratings of instruction. These were the summary statements and claims which guided this study in the selection and comparison of three types of response cues to determine their qualities of bias-proneness and rater reliability for the purpose of improving an existing student instructional rating scale.

CHAPTER III

DESIGN AND PROCEDURES

Introduction

This study was designed to test the effect of alternate response definitions on the leniency bias and rater reliability of student instructional rating form items. Three response formats were defined as (1) fixed alternative Likert cues (SA-SD), (2) fixed alternative evaluative cues (superior-inferior), and (3) multiple choice short descriptive cues. Leniency bias was measured by finding the closeness of each item mean to the midpoint of the rating scale. Since student ratings were overwhelmingly concentrated at the upper end of the scale, the format that gave the lowest mean was regarded as the least biased. In addition to finding the least biased and most reliable of these item types for student rating of instruction, the study tested two claims made in the literature:

a. Evaluative cues are more susceptible to response bias than other cues.

b. Fixed response alternatives are more susceptible to bias than descriptive multiple choice alternatives.

The study focussed on leniency bias in these tests because it was the kind of bias to which student rating forms were

shown to be susceptible, and because the intent of the study was to generate information that could be used to improve an existing instructional rating form.

In order to test the effect of alternate response definitions on leniency bias and rater reliability in a controlled setting, it was desirable to eliminate as many extraneous variables as possible from the comparison. This involved making the question stems as nearly alike as possible, manipulating only those response cue characteristics pertinent to the cue types, and insuring that the raters and rating situation were as nearly equivalent for the three forms as possible. The steps that were taken to develop the forms and test them in comparable situations are described in the sections entitled "Sample," "Instruments," and "Design." The sections entitled "Hypotheses" and "Analysis" restate the specific statements to be tested and describe the statistical methods utilized to determine the outcomes of the study.

Sample

Thirty-five instructors teaching courses with at least 30 students enrolled were asked to volunteer 20 minutes of class time within the last three weeks of Winter quarter 1973, to administer the instructional rating forms for the study. Twenty-five agreed to take part. Those that declined claimed too little time left in the quarter, or that they had to give the MSU form

that quarter. Two of the 25 participants had less than 25 students responding so were not included in the analysis. Table 3.1 reports the number of students responding to each form for each instructor.

Twenty participants taught undergraduate courses in the departments of Social Science (7), Humanities (6), Natural Science (6), and Education (1). Three taught masters-level courses in Education.

Instruments

Three instruments with machine-scorable answer sheets were compared in the study. The same 21 questions were asked on each instrument, with the wording not being changed any more than was necessary to accommodate the different response option types. The questions used were the first 21 questions on the MSU Student Instructional Rating Form. The remaining questions were student background questions and optional questions regarding rating of laboratory or recitation sections.

Instrument 1

Instrument 1 consisted of the unchanged first 21 questions of the Student Instructional Rating Scale currently in use at MSU. In the traditional Likert approach, it merely asked for the extent of agreement to statements, not for obviously normative evaluations.

Sample question:

(1) The instructor was enthusiastic when presenting course material
    1. Strongly agree
    2. Agree

TABLE 3.1.--Number of student raters responding to each form for each instructor.

Instructor	Likert	Evaluative	Descriptive	Total Class Size
1	12	14	15	41
2	13	11	13	37
3	13	11	12	36
4	10	12	13	35
5	11	11	12	34
6	27	24	17	61
7	16	17	16	49
8	19	17	13	49
9	9	8	8	25
10	20	20	17	57
11	12	11	13	36
12	22	22	24	68
13	18	14	18	50
14	13	16	15	44
15	14	14	13	41
16	11	14	14	39
17	21	25	20	66
18	17	17	15	49
19	11	12	14	37
20	14	15	16	45
21	9	8	10	27
22	18	20	20	58
23	15	15	15	45
TOTALS	345	348	343	N = 1,036

3. Neutral
4. Disagree
5. Strongly disagree

Instrument 2

Instrument 2 consisted of the same 21 questions as the first form, modified slightly to read smoothly with general, norm-referenced response options having "Average: typical of courses or instructors" as the mid-scale referent. In the evaluative format, it asked each rater to make a comparative judgment of the instructor relative to others in his experience.

Sample question:

(1) The instructor's enthusiasm when presenting course material
    1. Superior: exceptionally good course or instructor
    2. Above Average: better than the typical course or instructor
    3. Average: typical of courses or instructors
    4. Below Average: not as good as the typical
    5. Inferior: improvement definitely needed

Instrument 3

Instrument 3 consisted of exactly the same 21 question stems as Instrument 2 with descriptive graphic response options. The options were behavioral terms unique to each question derived by describing the ideal instructor, the average instructor, and the inferior instructor. The responses in this format contained more

information than the two types above since they were specific to one question only and did not have to be general enough to handle all.

Sample question:

(1) The instructor's enthusiasm when presenting course material

1	2	3	4	5
vibrant, stimulating		sometimes inspired		apathetic

Development of the Scales

The three instruments described above were developed from the original MSU form with 21 questions in Likert format. The question content in the three instruments was unchanged except as was necessary to make it read smoothly with the particular response cue format. The first instrument was in fact the original MSU form and so was unchanged in any way for the study.

The development of the second form was guided by the numerous examples of a "traditional" evaluative form found in the literature. The description, "Average: typical of courses or instructors" was selected as the midscale referent to reinforce the normal curve concept of average as the quality of achievement attained by the majority of the group. If the student raters were able to employ this definition accurately, most ratings would be within plus or minus one unit of this midpoint.

The development of the descriptive form was more detailed than the others since a different set of response cues was needed for each question. It was necessary to generate concise behavioral terms describing the ideal instructor, the average instructor, and the inferior instructor for each of the twenty-one questions. Some difficulty was experienced in generating descriptors of the average instructor which were not too positive or too negative sounding. Above all, the attempt was made to describe each type of instructor accurately. To avoid relying entirely on the experimentor's judgment in developing these descriptors, two forms of each item were developed and pretested. The pretest was conducted by administering the two forms to random halves of one class with 40 students. The object of this administration was to determine the less ambiguous descriptors in each pair. Since all students responding to a given item had had the same instruction and should ideally be expected to agree in their ratings, the variance of responses to each form was employed as a measure of ambiguity. If the amount of disagreement (variance in ratings) was greater for one item wording than for another wording, it was concluded that the wording was at fault. The item in each pair with the lesser variance was kept for the final scale. The items and their variance measures are reproduced in Table 3.2. (Figure 3.3 shows the final form used in the study.)

TABLE 3.2.--Pretest variances of two forms of each item of the descriptive scale.

Form	Item	Σx^2[a]
	1. The instructor's enthusiasm when presenting course material 1 2 3 4 5	
A	vibrant, stimulating — sometimes inspired — apathetic	7.8
B	animated, challenging — sometimes inspired — dull, bored	21.8
	2. The instructor's interest in teaching 1 2 3 4 5	
A	obviously enjoys — seems to enjoy — seems to dislike	10.5
B	imaginative, original — seems to enjoy — stereotyped, routine	14.5
	3. The instructor's use of examples or personal experiences to help get points across in class 1 2 3 4 5	
A	insightful — useful — inappropriate	19.2
B	very effective — sometimes helpful — waste of time	22.5
	4. The instructor's concern with whether the students learned the material 1 2 3 4 5	
A.	gave all possible help — helpful when asked — reluctant to help	11.0
B.	seemed to care a lot — seemed mildly concerned — seemed not to care	13.8

[a]Since variance is $\Sigma x^2/N$ and N = 20 in both groups, Σx^2 alone is calculated.

[b]Items 8, 16, and 18 are exceptions where, due to omitted responses, N's were unequal and variance was calculated.

Table 3.2.--Continued.

	5. Your interest in learning the course material					
	1	2	3	4	5	
A	avid		attentive		indifferent	13.2
B	eager		attentive		apathetic	20.2
	6. Your general attentiveness in class					
	1	2	3	4	5	
A	alert		observant		bored	24.8
B	attentive all the time		attentive most of time		rarely attentive	17.2
	7. This course as an intellectual challenge					
	1	2	3	4	5	
A	powerful		adequate		weak	14.5
B	stimulating		interesting		dull	20.5
	8. Your competence in this area due to this course					
	1	2	3	4	5	
A	much improved		sufficiently improved		doubtfully improved	.99[b]
B	increased greatly		increased moderately		increased little	.86[b]
	9. The amount of encouragement to students to express their opinions					
	1	2	3	4	5	
A	seeks out opinions		allows some time for		avoids discussion	16.4
B	rewards stating opinions		neutral toward stating		discourages stating	8.2

[b]Items 8, 16, and 18 are exceptions where, due to omitted responses, N's were unequal and variance was calculated.

Table 3.2.--Continued.

	10. The instructor's receptiveness to new ideas and others' viewpoints				
	1	2	3	4	5	
A	welcomes differences		usually tolerant		hostile to differences	22.0
B	seems to value them		allows them		seems to disdain them	31.2
	11. The student's opportunity to ask questions				
	1	2	3	4	5	
A	always available		sometimes available		never available	10.8
B	always available		frequently available		seldom available	13.0
	12. The instructor's stimulation of class discussion				
	1	2	3	4	5	
A	skillful		competent		awkward	11.0
B	occurred often		occurred frequently		occurred rarely	11.2
	13. The amount of material the instructor attempted to cover				
	1	2	3	4	5	
A	appropriate		somewhat pressured		too much material	25.0
B	ideal		reasonable		unreasonable	23.2
	14. The pace at which the instructor attempted to cover the material				
	1	2	3	4	5	
A	responsive to class tempo		rushed at times		always rushed	25.0
B[C]	responsive to class tempo		rushed at times		always rushed	25.8

[C]Unable to construct reasonable alternate for this item (see p. 62).

Table 3.2.--Continued.

		1	2	3	4	5	
	15. The homework assignments' contribution to your understanding of the course						
A		well worth time spent		mainly worth time spent		not worth time spent	20.0
B		great help		some help		no help	26.6
	16. The difficulty level of topics covered in assigned readings						
A		ideal		somewhat rigorous		too rigorous	.90[b]
B		appropriate to class		somewhat difficult		too difficult	1.60[b]
	17. The instructor's ability to relate course concepts in a systematic manner						
A		showed unity of topics		gave orderly presentation		made no effort to unify	16.6
B		showed topics' interrelation		gave orderly presentation		no relationship shown	21.0
	18. The organization of the course						
A		topics well arranged		could be improved		topics poorly arranged	.73[b]
B		unified and coherent		could be improved		disorganized	.92[b]
	19. The ease of taking notes on the instructor's presentation						
A		easy to take notes		possible with some effort		difficult to take notes	34.6
B		aided by instructor		normal situation		hampered by instructor	21.0

[b]Items 8, 16, and 18 are exceptions where, due to omitted responses, N's were unequal and variance was calculated.

Table 3.2.--Continued.

	20. The direction of the course	1	2	3	4	5	
A		clearly communicated		adequately communicated		poorly communicated	21.8
B		well defined		moderately well defined		not clearly defined	24.0
	21. Your general enjoyment of the class	1	2	3	4	5	
A		very enjoyable		enjoyable		distasteful	15.8
B		loved it		satisfied		hated it	27.2

[a]Since variance is $\Sigma x^2/N$ and N = 20 in both groups, Σx^2 alone is calculated.

[b]Items 8, 16, and 18 are exceptions where, due to omitted responses, N's were unequal and variance was calculated.

[c]Unable to construct reasonable alternate for this item (see p. 62).

In the process of writing the evaluative and descriptive response forms of the original Likert questions, it was discovered that four items concerning the topic of Course Demands (#13-16) were not parallel in scale format to the other items. For most items in the questionnaire, the first option on the scale was the highest rating an instructor could receive, but for these items the third option was the highest rating. This was due to the wording of the four questions such that the first option was a response of "too much" and the fifth option was a response of "too little." For example, in the question, "The instructor attempted to cover too much material. (SA-SD)," a response of "SA" meant "too much material" and a response of "SD" meant "too little material." The difficulty with this change in scale format was in the inability of the study to compare the mean ratings of the items where "3" was the highest rating to the mean ratings of their counterparts written in evaluative format, where "1" ("Superior") was the highest rating. The descriptive format could have been written to correspond to either scale, but no satisfactory transformation of all three scales was seen to be possible. This was not felt to be a condemnation of any one scale, but rather an unforseen difficulty in the study. It was concluded that the comparison of the remaining 17 items would give sufficient grounds to answer the hypotheses of the study, so the four questions were omitted from the analysis.

Figures 3.1, 3.2, and 3.3 are photographic reproductions of the machine-scorable forms administered in the study.

Design

Three instructional rating forms differing primarily in response cue format were developed and administered to randomly equivalent thirds of each class of 23 instructors.

Each instructor was given a packet containing the three forms arranged alternately so that (assuming a random start) each form would be automatically distributed to random thirds of the class. Each student received one form. Directions were given to the instructors to administer the forms just as they have administered the instructional rating form in the past. Differences in administration, if present, were considered a legitimate potential source of variance in instructors. The instructors were told and could pass on to their students that a new form was being tried out. But neither they nor their students were informed of the research hypotheses regarding leniency or reliability. The answer sheets were collected and machine-scored and the data punched onto cards. Instructors were assured of anonymity of results.

Generalizability of Results

The nonrandom selection of instructors did not affect the comparison of rating forms to each other since

Figure 3.1.--The MSU student instructional rating form.

MICHIGAN STATE UNIVERSITY

STUDENT INSTRUCTIONAL RATING SYSTEM FORM

SA - if you strongly agree with the statement
A - if you agree with the statement
N - if you neither agree nor disagree
D - if you disagree with the statement
SD - if you strongly disagree with the statement

Please omit any of the items which do not pertain to the course that you are rating. For example, if you have had no homework assignments in this course omit (leave blank) those items pertaining to homework. With a pencil respond to the items using the KEY.

	KEY	SA	A	N	D	SD
1. The instructor was enthusiastic when presenting course material.	1.	SA	A	N	D	SD
2. The instructor seemed to be interested in teaching.	2.	SA	A	N	D	SD
3. The instructor's use of examples or personal experiences helped to get points across in class.	3.	SA	A	N	D	SD
4. The instructor seemed to be concerned with whether the students learned the material.	4.	SA	A	N	D	SD
5. You were interested in learning the course material.	5.	SA	A	N	D	SD
6. You were generally attentive in class.	6.	SA	A	N	D	SD
7. You felt that this course challenged you intellectually.	7.	SA	A	N	D	SD
8. You have become more competent in this area due to this course.	8.	SA	A	N	D	SD
9. The instructor encouraged students to express opinions.	9.	SA	A	N	D	SD
10. The instructor appeared receptive to new ideas and others' viewpoints.	10.	SA	A	N	D	SD
11. The student had an opportunity to ask questions.	11.	SA	A	N	D	SD
12. The instructor generally stimulated class discussion.	12.	SA	A	N	D	SD
13. The instructor attempted to cover too much material.	13.	SA	A	N	D	SD
14. The instructor generally presented the material too rapidly.	14.	SA	A	N	D	SD
15. The homework assignments were too time consuming relative to their contribution to your understanding of the course material.	15.	SA	A	N	D	SD
16. You generally found the coverage of topics in the assigned readings too difficult.	16.	SA	A	N	D	SD
17. The instructor appeared to relate the course concepts in a systematic manner.	17.	SA	A	N	D	SD
18. The course was well organized.	18.	SA	A	N	D	SD
19. The instructor's class presentations made for easy note taking.	19.	SA	A	N	D	SD
20. The direction of the course was adequately outlined.	20.	SA	A	N	D	SD
21. You generally enjoyed going to class.	21.	SA	A	N	D	SD
22.	22.	SA	A	N	D	SD
23. Instructor may insert three (3) items in these spaces.	23.	SA	A	N	D	SD
24.	24.	SA	A	N	D	SD

STUDENT BACKGROUND: Select the most appropriate alternative.

25. Was this course required in your degree program?	25.	yes	no			
26. Was this course recommended to you by another student?	26.	yes	no			
27. What is your overall GPA? (a) 1.9 or less (b) 2.0-2.2 (c) 2.3-2.7 (d) 2.8-3.3 (e) 3.4-4.5	27.	a	b	c	d	e
28. How many other courses have you had in this department? (a) none (b) 1-2 (c) 3-4 (d) 5-6 (e) 7 or more	28.	a	b	c	d	e
29. Instructor may insert two (2) items in this space.	29.	a	b	c	d	e
30.	30.	a	b	c	d	e

DO NOT WRITE BELOW THIS LINE UNLESS THIS COURSE HAS LABORATORY OR RECITATION SECTIONS

LABORATORY or RECITATION: (fill in your recitation or lab number at the bottom)

31. The laboratory or recitation instructor clarified lecture material.	31.	SA	A	N	D	SD
32. The laboratory or recitation instructor adequately prepared you for the material covered in his section.	32.	SA	A	N	D	SD
33. You generally found the laboratories or recitations interesting.	33.	SA	A	N	D	SD
34. Instructor may insert two (2) items in this space.	34.	SA	A	N	D	SD
35.	35.	SA	A	N	D	SD

IMPORTANT

WRITE and **MARK** in the boxes to the right your recitation or laboratory section number. Section number 1 would be written and marked 001; section number 15 would be written and marked 015. If you do not have a recitation or lab section leave this area blank.

RECITATION OR LABORATORY SECTION NUMBER

1.	0	1	2	3	4	5	6	7	8	9
2.	0	1	2	3	4	5	6	7	8	9
3.	0	1	2	3	4	5	6	7	8	9

Michigan State University Printing Service

Figure 3.2.--The experimental evaluative form.

STUDENT INSTRUCTIONAL RATING FORM--X5

For each item, respond by marking the number in the key that corresponds to the closest description of your instructor or your course.

PLEASE NOTE CHANGES IN THE KEY

1-Superior: Exceptionally good course or instructor
2-Above Average: better than the typical course or instructor
3-Average: typical of courses or instructors
4-Below Average: not as good as the typical
5-Inferior: one of the worst

1. The instructor's enthusiasm when presenting course material ------- 1
2. The instructor's apparent interest in teaching ------- 2
3. The instructor's use of examples or personal experiences to help get points across in class ------- 3
4. The instructor's concern with whether the students learned the material ------- 4
5. Your interest in learning the course material ------- 5
6. Your general attentiveness in class ------- 6
7. This course as an intellectual challenge ------- 7
8. This course's ability to improve your competence in this area ----- 8
9. The amount of encouragement to students to express opinions ------- 9
10. The instructor's receptiveness to new ideas and others' viewpoints- 10
11. The student's opportunity to ask questions ------- 11
12. The instructor's stimulation of class discussion ------- 12
13. The appropriateness of the amount of material the instructor attempted to cover ------- 13
14. The appropriateness of the pace at which the instructor attempted to cover the material ------- 14
15. The homework assignments' contribution to your understanding of the course material relative to the amount of time required --- 15
16. The appropriateness of the difficulty level of the coverage of topics in the assigned readings ------- 16
17. The instructor's ability to relate the course concepts in a systematic manner ------- 17
18. The course organization ------- 18
19. The ease of taking notes on the instructor's presentation -------- 19
20. The adequacy of the outlined direction of the course ------- 20
21. Your general enjoyment of the class ------- 21

Michigan State University Printing Service

Figure 3.3.--The experimental descriptive graphic form.

STUDENT INSTRUCTIONAL RATING FORM--X4

For each item, respond by marking the number in the key that corresponds to the closest description of your instructor on each continuum. If he or she fits the description under 3, mark 3. If he or she is somewhere between the descriptions under 3 and 1, mark 2. There are 5 choices for each question.

1. The instructor's enthusiasm when presenting course material

1	2	3	4	5	
vibrant, stimulating		sometimes inspired		apathetic	1

2. The instructor's interest in teaching

1	2	3	4	5	
obviously enjoys		seems to enjoy		seems to dislike	2

3. The instructor's use of examples or personal experiences to help get points across in class

1	2	3	4	5	
insightful		useful		inappropriate	3

4. The instructor's concern with whether the students learned the material

1	2	3	4	5	
gave all possible help		helpful when asked		reluctant to help	4

5. Your interest in learning the course material

1	2	3	4	5	
avid		attentive		indifferent	5

6. Your general attentiveness in class

1	2	3	4	5	
attentive all the time		attentive part of time		rarely attentive	6

7. This course as an intellectual challenge

1	2	3	4	5	
powerful		adequate		weak	7

8. Your competence in this area due to this course

1	2	3	4	5	
increased greatly		increased moderately		increased little	8

9. The amount of encouragement to students to express their opinions

1	2	3	4	5	
rewards stating opinions		neutral toward stating		discourages stating	9

10. The instructor's receptiveness to new ideas and others viewpoints

1	2	3	4	5	
welcomes differences		usually tolerant		hostile to differences	10

11. The student's opportunity to ask questions

1	2	3	4	5	
always available		sometimes available		never available	11

12. The instructor's stimulation of class discussion

1	2	3	4	5	
skillful		competent		awkward	12

13. The amount of material the instructor attempted to cover

1	2	3	4	5	
ideal		reasonable		unreasonable	13

14. The pace at which the instructor attempted to cover the material

1	2	3	4	5	
responsive to class tempo		rushed at times		always rushed	14

15. The homework assignments contribution to your understanding of the course

1	2	3	4	5	
well worth time spent		mainly worth time spent		not worth time spent	15

16. The difficulty level of topics covered in assigned readings

1	2	3	4	5	
ideal		somewhat rigorous		too rigorous	16

17. The instructor's ability to relate course concepts in a systematic manner

1	2	3	4	5	
showed unity of topics		gave orderly presentation		made no effort to unify	17

18. The organization of the course

1	2	3	4	5	
topics well arranged		could be improved		topics poorly arranged	18

19. The ease of taking notes on the instructor's presentation

1	2	3	4	5	
aided by instructor		normal situation		hampered by instructor	19

20. The direction of the course

1	2	3	4	5	
clearly communicated		adequately communicated		poorly communicated	20

21. Your general enjoyment of the class

1	2	3	4	5	
very enjoyable		enjoyable		distasteful	21

O-860.1

Michigan State University Printing Service

all instructors were rated with all forms by randomly equivalent groups of students. The fact that the instructors were volunteers did affect the ability to compare their mean ratings and rater reliabilities to other means and reliabilities reported in the literature.

Hypotheses

Null Hypotheses

H_1. There are no differences in mean ratings of instructors between items with Likert, evaluative, and descriptive response cue formats.

H_2. There are no differences in rater reliabilities between items with Likert, evaluative, and descriptive response cue formats.

Alternative Hypotheses

H_{1a}. The mean ratings with the evaluative format will be significantly more lenient than the mean ratings for the Likert and descriptive formats.

H_{1b}. The mean ratings with descriptive cue formats will be significantly less lenient than the mean ratings for Likert and evaluative cue formats.

H_{2a}. The descriptive response cue format will have significantly more rater reliability than the Likert or evaluative cue formats.

Analysis

The hypothesis of no differences in mean ratings was tested with a two-way multivariate analysis of variance design, instructor by treatment, where the response cue formats were the three treatments and the seventeen usable items were seventeen dependent variables. The data were first tested for the presence of interaction between instructors and treatments. A nonsignificant interaction would allow an overall F-test, $\alpha = .05$ for the main effect of treatment to determine whether the item means of any format were significantly different from the item means of any other format over all the items. Individual item F-values were also inspected to determine sources of variance with the understanding that lack of independence among the items prevents each individual F from having a known constant error. Scheffe post hoc analyses tested alternate hypotheses H_{1a} and H_{1b}. A packaged computer program written by Jeremy Finn was available to do the multivariate analysis of variance.

The hypothesis of no differences in rater reliabilities was tested by comparing confidence intervals about the reliability estimates for each format of each item. Overlap of confidence intervals would indicate no significant differences in rater reliabilities with the probability of no Type 1 error being $(1 - \alpha)^3$. If the number of items with non-overlapping intervals was greater

than chance, the null hypothesis was rejected. (Fisher's r to z transformation was not used because the rater reliability estimate for this rating situation includes between-raters variance in the error term whereas the Pearson reliability estimate excludes it. This would likely make the distribution of this coefficient different from the coefficient on which Fisher based his r to z transformation.)

Since the coefficient used to calculate the rater reliabilities was the intraclass rater reliability coefficient written in analysis of variance terms, it was possible to use the Finn MANOVA program to find the necessary components to generate the reliability estimates by hand. The necessary mean squares were obtained from a one-way raters-nested-within-instructors design with 17 dependent variables (the items) within each treatment group. Thus, since there were three treatment groups, three separate MANOVA's were necessary to generate the data for the rater reliability estimates.

The rater reliabilities and the confidence intervals were generated and tested by hand. The formula for the reliability of one average rater and the Spearman-Brown formula for the reliability of the average of k raters are:

$$r_{11} = \frac{MSB - MSE}{MSB + (k_o - 1)MSE} \quad \text{and} \quad r_{kk} = \frac{k\, r_{11}}{1 + (k-1)r_{11}}$$

where

MSB = mean square between instructors

MSE = mean square within instructors

k_o = average number of ratings per instructor in the sample

$$k_o = \frac{1}{n - 1}\left[\Sigma k_i - \frac{\Sigma k_i^2}{\Sigma k_i}\right]$$

n = number of instructors

k_i = number of ratings of each instructor

r_{11} = reliability of one average rater

r_{kk} = reliability of an average of k ratings

k = number of ratings for which a prediction of reliability is desired

The estimate of precision, as detailed by Lindquist (1953), is found by determining the upper and lower bounds of F and substituting them into the formula for r_{11}, where r_{11} is rewritten as $r_{11} = \frac{F_o - 1}{F_o + (k_o - 1)}$. The $(100 - 2\alpha)\%$ confidence interval for r_{11} becomes:

$$\frac{F_L - 1}{F_L + (k_o - 1)} < r_{11} < \frac{F_U - 1}{F_U + (k_o - 1)}$$

where

$F_L = F_o / F(\alpha)$, $\quad F_o = \frac{MSB}{MSE}$ in the sample

$F_U = F_o \cdot F_{(\alpha)}$

Figure 3.4 illustrates the design of the multivariate analysis of variance to be conducted for the tests of both hypotheses.

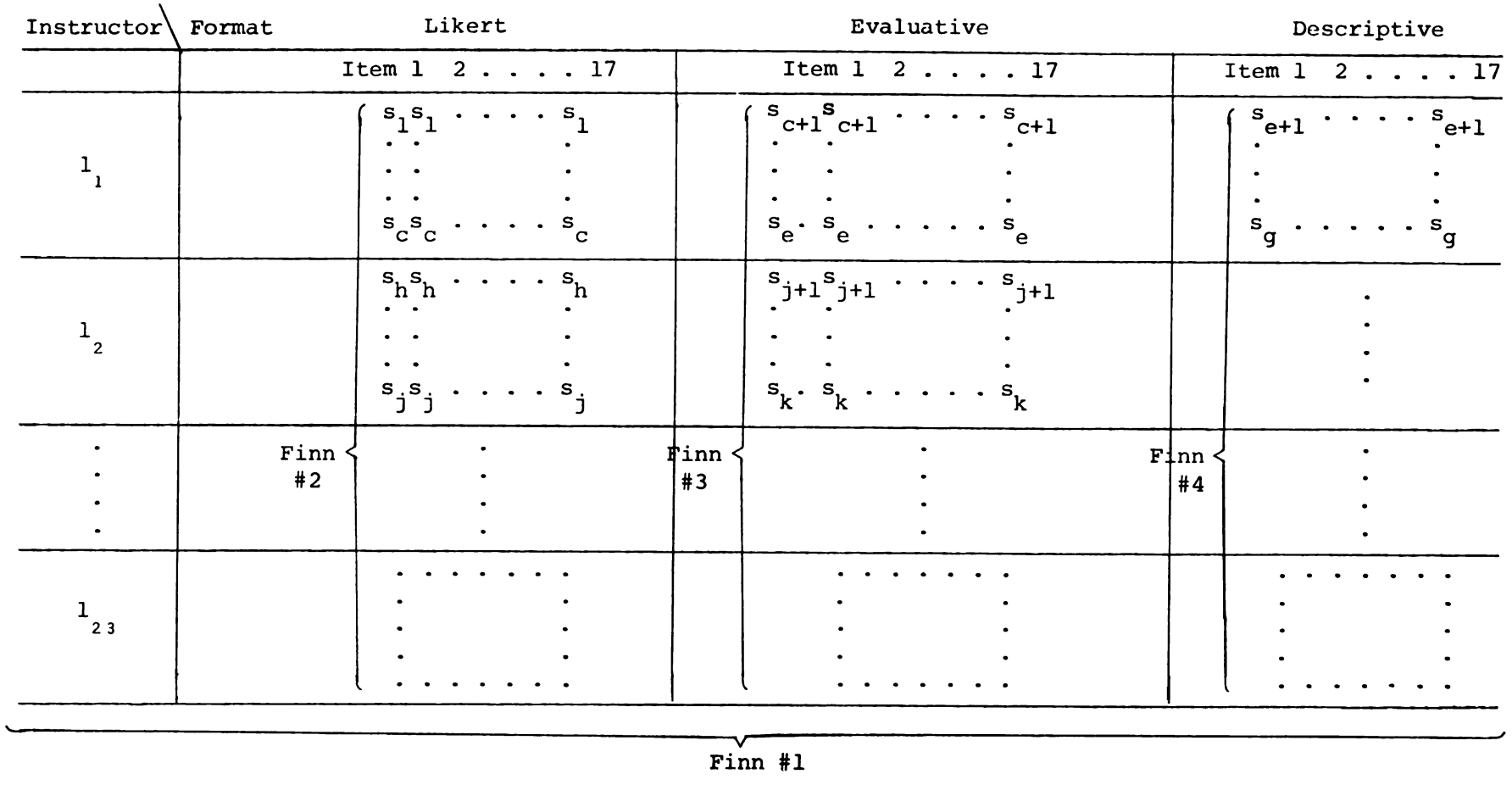

Figure 3.4.--Design of the experiment.

Summary

Three instructional rating forms differing primarily in response cue format were developed and administered to random thirds of the classes of 23 instructors. The three response cue formats were defined as (1) fixed alternative Likert cues (SA-SD), (2) fixed alternative evaluative cues (superior-inferior), and (3) multiple choice short descriptive cues. The questions used were the first 21 questions of the MSU student instructional rating scale. The questions were in Likert format on the original scale and were used unchanged. The question wording was altered slightly on the other two forms to accommodate the evaluative and descriptive response cues. The descriptive response cues were pre-tested to determine the least ambiguous descriptors for the final form.

The study was designed to test the effect of the alternate response definitions on the leniency bias and rater reliability of the three forms. Leniency bias was measured by finding the closeness of each item mean to the midpoint of the rating scale. Since student ratings were overwhelmingly concentrated at the upper end of the scale, the format that gave the lowest mean was regarded as the least biased. In addition to finding the least biased and most reliable response cue format for student rating of instruction, the study tested two claims made in the literature as they applied to the bias of lenient responding.

a. Evaluative cues are more susceptible to bias than other cues.

b. Fixed response alternatives are more susceptible to bias than descriptive multiple choice alternatives.

The test was limited to the question of leniency bias since this was the major problem with student ratings of instruction.

The hypothesis of no differences in mean ratings was tested with a two-way multivariate analysis of variance design, instructor by treatment, where the response cue formats were the three treatments and the 17 usable items were 17 dependent variables. Scheffe post hoc analyses tested alternate hypotheses that the evaluative format would produce the most lenient items, the Likert format the next most lenient, and the descriptive format the least lenient.

The hypothesis of no differences in rater reliabilities was tested by comparing confidence intervals about the reliability estimate for each format of each item. Non-overlapping confidence intervals would indicate significant differences in rater reliabilities.

CHAPTER IV

RESULTS

Introduction

The study was designed to test the leniency bias-proneness and rater reliability of three response cue formats. The major reason for conducting the study was to improve an existing Likert-type student instructional rating scale. Since the content of the scale was well established in its creation, the study was focussed on manipulating the response options to reduce the amount of lenient responding present with the existing scale. Two alternative response definitions were chosen to compare with the existing Likert format response definitions. The descriptive graphic format was chosen as the most often recommended format for reducing leniency bias. It was hypothesized that this format would produce the least lenient responses from student raters of instruction. The evaluative format was chosen as a second alternative for purposes of contrast because it was claimed to be the most bias-prone response format. It was hypothesized that the evaluative format would produce the most lenient responses, and the descriptive the least lenient responses.

It was also hypothesized that the least lenient response cue format would prove to have the greatest rater reliability.

A concurrent purpose of the study was to test two claims made in the literature concerning the bias-proneness of certain response definitions. The testing of the claims as they applied to the bias of lenient responding was compatible with the major objective of the study. The claims to be tested were:

1. Evaluative cues are more susceptible to response bias than other cues.

2. Fixed response alternatives are more susceptible to bias than descriptive multiple choice alternatives.

If both of these claims were to hold true for leniency bias, the hypothesized order of response cue types, from most to least lenient would be, fixed alternative evaluative cues, fixed alternative Likert cues, and multiple choice short descriptive cues.

To conduct the study, three instructional rating forms differing primarily in response cue format were developed and administered to random thirds of the classes of 23 instructors. Leniency bias was measured by finding the closeness of each item mean to the midpoint of the rating scale. Since student ratings were overwhelmingly concentrated at the upper end of the scale, the format that gave the lowest mean was regarded as the least biased.

The hypothesis of no differences in mean ratings (leniency bias) was tested with a two-way multivariate analysis of variance design, instructor by treatment, where the response cue formats were the three treatments and the 17 usable items were 17 dependent variables. Scheffe post hoc analyses tested alternate hypotheses that the evaluative format would produce the most lenient items, the Likert format the next most lenient, and the descriptive format the least lenient.

The hypothesis of no differences in rater reliabilities was tested by comparing confidence intervals about the reliability estimate for each item. Non-overlapping confidence intervals would indicate significant differences in rater reliabilities.

The following sections present the results concerning leniency bias and the results concerning rater reliability of the three instructional rating forms with alternate response definitions.

Results Concerning Leniency Bias

The test of Hypothesis 1 was carried out by the test of the main effect of treatment in the two-way, instructor by treatment, analysis of variance design. Hypothesis 1 was stated,

> 1. There are no differences in mean ratings of instructors between items with Likert, evaluative, and descriptive response cue formats.

A significant treatment effect, $F = 10.40$ $p < .0001$, indicated that the mean ratings of the 23 instructors were different with the three instructional rating forms, even though the instructors being rated were the same for each form and the groups of students rating them with the different forms were assumed to be randomly equivalent. A possible complicating factor in such a design, which had to be tested also before a clear result could be established, was the interaction effect of instructor with treatment. Interaction would have occurred if some instructors received their least lenient ratings with one response format while others received their least lenient ratings with another format. It would not have been possible to establish a clear format effect if such an interaction were present in the results, since the order of most and least lenient response formats would have depended on the instructor being rated. Fortunately, interaction effects were not found to be significant in the study results, indicating that the three response cue formats produced clear differences in lenient responding for all instructors (Table 4.1).

After establishing that differences in mean ratings between the response formats existed, steps were taken to determine which items were contributing to the significant overall difference in means and to determine which formats were producing the most and least lenient means.

TABLE 4.1.--F-ratios, instructor by treatment MANOVA.

Effect	F	DF_1	DF_2	p less than
Instructor	5.1012	374	12891.01	.0001
Treatment	10.4026	34	1902.00	.0001
Interaction	1.0128	748	15184.85	.3989

In the multivariate analysis of variance, each item was a contributing dependent variable for which independent tests of significance were carried out in addition to the overall test which led to the rejection of the hypothesis of no differences in item means. Inspection of the individual item F tests of treatment effect indicated that most items were contributing to the significant overall difference in item means found between response cue formats. Such an inspection was used to indicate sources of differences, though lack of independence among the items prevented each individual F from having a known constant error. The large number of significant item F's did indicate that the effect of response cue format was present with most items and not limited to a few (Table 4.2).

To summarize the results thus far, the null hypothesis of no differences in item means was rejected, establishing that there were significant differences in leniency between the three instructional rating forms.

TABLE 4.2.--Univariate F tests, each dependent variable (each item).

Variable	Mean Square	Univariate F	p less than
Item 1: I-enthusiasm[a]	10.1345	20.4571	.0001
Item 2: I-interest	5.6580	10.8606	.0001
Item 3: I-examples	7.8597	9.0612	.0002
Item 4: I-concern	7.1532	8.4603	.0003
Item 5: S-interest[b]	20.7714	21.7417	.0001
Item 6: S-attention	22.6673	31.6897	.0001
Item 7: S-challenge	11.5925	11.6609	.0001
Item 8: S-competence	8.1133	7.9277	.0004
Item 9: opinions	4.4629	6.0640	.0025
Item 10: new ideas	11.0392	14.9030	.0001
Item 11: questions	27.4122	46.3598	.0001
Item 12: discussion	7.2542	8.5008	.0003
Item 17: unity of topics	2.4900	2.8058	.0610
Item 18: organization	11.2256	13.7279	.0001
Item 19: note-taking	3.9678	3.3147	.0368
Item 20: course outline	9.1462	10.4170	.0001
Item 21: enjoyment	3.4328	2.8815	.0566

[a]"I" stands for "Instructor."

[b]"S" stands for "Student."

Scheffe post hoc analysis of the directions of the differences between individual item means indicated that the predicted directions of differences were only partially correct.

The alternate hypotheses were:

H_{1a}. The mean ratings with the evaluative format will be significantly more lenient than the mean ratings for the Likert and descriptive formats.

H_{1b}. The mean ratings with descriptive cue formats will be significantly less lenient than the mean ratings for Likert and evaluative cue formats.

In order for both alternate hypotheses to be correct, item means in the three response cue formats would

have had to have been ordered so that the evaluative format produced the most lenient items, the Likert the next most lenient, and the descriptive the least lenient.

A contrast of Likert and descriptive item formats showed that the descriptive format was in fact less lenient than the Likert format as predicted. Seven descriptive item means were significantly less lenient than their Likert format counterparts, one was significantly different in the opposite direction, and the rest were not significantly different. The probability that seven out of 17 tests would be significant by chance alone, $\alpha = .05$, is less than .001 assuming independent tests, so it was concluded that the descriptive format cues were significantly less lenient than the Likert format response cues (Sakoda et al., 1954).

The ordering of the item means in the study data differed from the ordering predicted by the alternate hypotheses in that the evaluative format produced completely opposite results to those predicted. Instead of being the most lenient response format, it was found to be the least lenient of all the formats. Post hoc analysis showed that the evaluative format was significantly less lenient than the Likert format in 15 out of 17 items, and that it was significantly less lenient than the descriptive format in 10 out of 17 items. Since the majority of items produced this effect, it was concluded that the evaluative

format was the least lenient format in this study. The contrasts of item means are presented in Table 4.3. The combined results of the tests of significance of the contrasts between all item means are represented in the final column of the table by orderings with "less than" signs depicting significant differences and "equal" signs depicting non-significant differences. The items are grouped according to the factors established when the original scale was constructed in order to compare the performances of items on the same topics. Due to the

TABLE 4.3.--Contrasts of item means.

Item	L-E	D-E	L-D	Order
Factor 1: Instructor Involvement				
1: I-enthusiasm	-.29*	.02	-.31*	L<D=E
2: I-interest	-.25*	-.15*	-.10*	L<D<E
3: I-examples	-.13	.17*	-.30*	L=E<D
4: I-concern	-.28*	-.18*	-.10	L=D<E
Factor 2: Student Interest				
5: S-interest	-.38*	.08	-.46*	L<D=E
6: S-attention	-.45*	-.43*	-.02	L=D<E
7: S-challenge	-.36*	-.22*	-.15	L=D<E
8: S-competence	-.26*	.01	-.27*	L<D=E
Factor 3: Student-Instructor Interaction				
9: opinions	-.19*	-.20*	.01	L=D<E
10: new ideas	-.26*	-.34*	.07	L=D<E
11: questions	-.36*	-.56*	.20*	D<L<E
12: discussion	-.24*	.02	-.26*	L<D=E
Factor 4: Course Organization				
17: unity of topics	-.17*	-.07	-.09	L<E
18: organization	-.22*	-.36*	.13	L=D<E
19: note-taking	-.17*	-.20*	.02	L=D<E
20: course outline	-.32*	-.20*	-.13	L=D<E
21: enjoyment	-.10	.10	-.20*	L<D

*Significant, $\alpha = .05$.
L = Likert, E = evaluative, D = descriptive.

machine-scoring of the items, the lowest numbers were given to the most lenient responses, hence the order of results will seem to be reversed. The greatest numbers are really the least lenient responses in this system.

The most common ordering of means, occurring with eight items, was L=D<E. This ordering summarizes the significant differences L<E and D<E, and the nonsignificant difference between L and D, L=D. It means that the evaluative format produced less lenient responses than the other two formats but that the differences between the Likert and descriptive formats were not distinguishable in these items. The second most common ordering, occurring with four items, was L<D=E. Here, the Likert format is clearly the most lenient format since L<D and L<E, but for these items, the differences between D and E are not distinguishable. It appears from these combined results that the only ordering compatible with all 12 items would be L<D<E, where the Likert format is most lenient, the descriptive format the next most lenient, and the evaluative format the least lenient. (In fact, this ordering is compatible with the results of all but two items tested.) The nonsignificant differences between descriptive and Likert means in some items and descriptive and evaluative means in other items appeared to be due to the variability of the descriptive format itself, since the other two formats maintained a relatively constant distance between each other.

A graph of item means for the three response cue formats (Figure 4.1) illustrates the results discussed above. The evaluative format is shown to be consistently less lenient than the Likert format, while the differences between the descriptive item format and the other formats were shown to be not as consistent nor as great. But the graph also showed that the descriptive item means were less lenient than the Likert format item means for all but three items, further supporting the compatibility of the ordering L<D<E with the results. The table of item means (Table 4.4) shows the actual values obtained, with the significance of the orderings restated in the last column.

It was apparent from the graph that the three formats all produced a similar profile of high and low item ratings for the instructor group. This implied that the question stems were essentially the same; i.e., that the traits rated were probably the same for each format. However, for some items, the descriptive format appeared to magnify the differences between item ratings producing higher high ratings and lower low ratings for the various items.

Inspection of the table of means showed that while their orders supported the conclusion that Likert was most lenient, descriptive next most lenient, and evaluative least lenient, their values in this study were still within the upper half of the five-point scale for all response

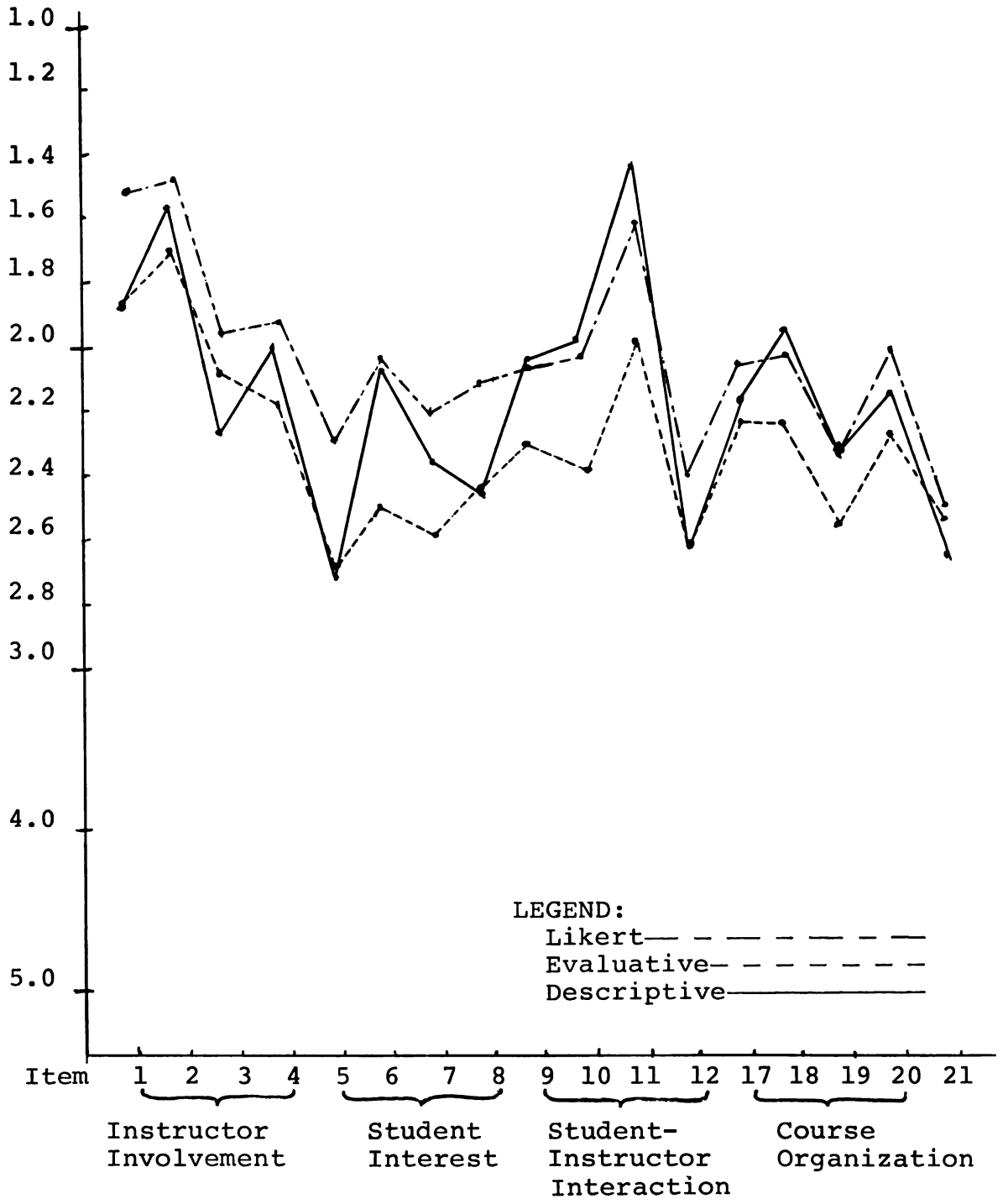

Figure 4.1.--Graph of item means for the three response cue formats.

TABLE 4.4.--Table of item means for the three response cue formats.

Item		Likert	Evaluative	Descriptive	Order
Factor 1: Instructor Involvement					
1:	I-enthusiasm[a]	1.56	1.85	1.86	L<D=E
2:	I-interest	1.49	1.74	1.59	L<D<E
3:	I-examples	1.99	2.12	2.29	L=E<D
4:	I-concern	1.93	2.20	2.01	L=D<E
Factor 2: Student Interest					
5:	S-interest[b]	2.29	2.67	2.74	L<D=E
6:	S-attention	2.06	2.51	2.08	L=D<E
7:	S-challenge	2.24	2.60	2.38	L=D<E
8:	S-competence	2.19	2.45	2.46	L<D=E
Factor 3: Student-Instructor Interaction					
9:	opinions	2.16	2.35	2.13	L=D<E
10:	new ideas	2.13	2.40	2.05	L=D<E
11:	questions	1.66	2.01	1.44	D<L<E
12:	discussion	2.43	2.65	2.65	L<D=E
Factor 4: Course Organization					
17:	unity of topics	2.08	2.25	2.18	L<E
18:	organization	2.06	2.28	1.93	L=D<E
19:	note-taking	2.39	2.57	2.39	L=D<E
20:	course outline	2.01	2.33	2.14	L=D<E
21:	enjoyment	2.48	2.57	2.67	L<D

[a]I = instructor. [b]S = student.

formats. This could have been an indication that stronger measures than changing response cue format would be required to reduce lenient responding, but it was also expected that the values were slightly inflated due to the quality of instructors that were likely to volunteer to be evaluated.

Each mean discussed above was the average of the ratings of all 23 instructors on a particular item and format. It was a summary value representing the item means of 23 individual instructors. If each item mean in the

table above were replaced by the 23 individual instructor means contributing to it, a very large table with 23 x 17 means in each of the three format categories would result. The table is not reproduced here, but it was created and did provide the data to determine the range of mean ratings produced by each of the three formats. A summary of these results is presented in Table 4.5.

The major considerations in the inspection of the range of item means produced by each scale format were the number and percentage of extremely lenient means and the number and percentage of non-lenient means. Extremely lenient means were defined as means of 1.0 (perfect) to 1.6. Non-lenient means were defined as means greater than or equal to the scale midpoint of 3.0. Indications of the breadth and skewness of variation were obtained from these figures.

The results of this inspection were consistent with the results previously obtained in the post hoc analysis.

TABLE 4.5.--Number and percentage of extreme instructor means.

Format	Less than 1.6		Greater than 2.9		Total
	Number	Percentage[a]	Number	Percentage	Number
Likert	54	13.8	17	4.3	71
Evaluative	33	8.4	40	10.2	73
Descriptive	48	12.3	35	9.0	83

[a]Percentage of 391 total means in each treatment.

The Likert format produced the most lenient means and the least non-lenient means yielding the most leniently skewed distribution. The evaluative format produced the least lenient means and the most non-lenient means yielding the least leniently skewed distribution. The descriptive format produced a moderately high number of both lenient and non-lenient means yielding the most variable distribution. The descriptive format produced a greater total number of means in the extreme categories than either of the other two formats.

These results are further illustrated in the percentage of total means in each extreme category. The Likert format is clearly the most skewed, having 13.8% of total means in the most lenient category and 4.3% of total means in the least lenient category. The evaluative format is more evenly balanced with 8.4% in the upper range and 10.2% in the lower range. The descriptive format is confirmed to be most variable, with nearly as great a percentage of very lenient means as the Likert format and nearly as great a percentage of non-lenient means as the evaluative format. None of the formats had a large percentage of means in the lower half of the scale.

Inspection of the graph and of the questions themselves indicated that the topic of the questions (the particular factor) was sometimes related to the type of response format that worked best (least leniently) with it.

The majority of questions concerning the factor of Course Organization were answered least leniently in the evaluative format with no differences being found between the Likert and descriptive item types. Similarly, the majority of questions concerning Student-Instructor Interaction were answered least leniently in the evaluative format, though the performance of L and D items on this topic was more variable (L=D, L=D, D<L, L<D). Less consistent relationships between item type and topic were obtained for the topics of Instructor Involvement and Student Interest. For half of these items, the evaluative format was clearly the least lenient, but for the other half, the descriptive format showed the most promise, equalling or surpassing the evaluative format means. These results might indicate that while evaluative items were the best overall rating measure, descriptive items were as good for rating people. No topics contained questions that were answered least leniently by Likert format items.

The claimed superiority of multiple choice over fixed alternative response cue formats in reducing bias was not substantiated by the results of this study of leniency. The multiple choice descriptive items were found to be less lenient than fixed alternative evaluative items. Fixed alternative item types produced both most lenient and least lenient responding in the study, contradicting the claimed superiority of multiple choice items in reducing bias.

In summary, it was found that the leniency of the three response cue formats was only partially predicted by the alternate hypotheses. The evaluative format was not found to be the most lenient in this study, but rather the least lenient. The descriptive format produced less lenient responses than the Likert format as predicted, but it was more variable in its influence on lenient responding than the evaluative format. The Likert format was found to be the most often prone to leniency bias. Inspection of the range of instructor means produced by each format corroborated the findings concerning the leniency of the various formats. The evaluative format produced the smallest percentage of extremely lenient means while the Likert format produced the greatest percentage of them. It was noted that while there were statistically significant differences in lenient responding between the formats, the majority of instructor means in all formats was concentrated in the upper half of the scale. This was thought to be due in part to the higher quality of instructors who would volunteer to be evaluated.

Inspection of the graph of item means and of the questions themselves indicated that the topic of questions (the particular factor) was sometimes related to the type of response format that worked best (least leniently) with it. Although the majority of the questions in all topics were answered least leniently in evaluative format, it

was most successful with the topics of Course Organization and Student-Instructor Interaction. The descriptive format showed the most promise with questions concerning Instructor Involvement and Student Interest, equalling or surpassing the evaluative format means in half of the items. These results seemed to indicate that while evaluative items were the best overall rating measure, descriptive items were as good for rating people.

The claim concerning multiple choice vs. fixed alternative cues was not substantiated by the results of this study. Fixed alternative response formats were found to produce both the most lenient and the least lenient responding. The multiple choice descriptive response format was moderately successful in reducing lenient responding, but the fixed alternative evaluative item type was consistently more successful in reducing lenient responding.

Results Concerning Rater Reliability

The hypothesis of no difference in rater reliabilities between items with Likert, evaluative, and descriptive response cue formats in student rating of instruction was not rejected. The method of comparing confidence intervals for each item produced overlapping intervals for all response cue formats (Table of confidence intervals, Appendix).

Hypothesis 2 was stated,

H_2. There are no differences in rater reliabilities between items with Likert, evaluative, and descriptive response cue formats.

The item reliabilities for a single average rater (Table 4.6) were inspected to determine whether there was a prevalent order of rater reliabilities according to response format despite the lack of significant differences found by comparison of confidence intervals. It was hoped

TABLE 4.6. Item reliabilities for a single average rater.

Item	Likert	Evaluative	Descriptive	Order[a]
Factor 1: Instructor Involvement				
1: I-enthusiasm[b]	.15	.17	.27	L<E<D
2: I-interest	.10	.13	.18	L<E<D
3: I-examples	.11	.13	.21	L<E<D
4: I-concern	.10	.14	.14	L<E=D
Factor 2: Student Interest				
5: S-interest[c]	.11	.07	.09	E<D<L
6: S-attention	.06	.09	.03	D<L<E
7: S-challenge	.13	.10	.13	E<L=D
8: S-competence	.01	.05	.07	L<E<D
Factor 3: Student Instructor Interaction				
9: opinions	.34	.26	.23	D<E<L
10: new ideas	.23	.17	.16	D<E<L
11: questions	.15	.15	.08	D<E=L
12: discussion	.30	.30	.21	D<E=L
Factor 4: Course Organization				
17: unity of topics	.07	.08	.08	L<E=D
18: organization	.08	.07	.08	E<L=D
19: note-taking	.15	.14	.12	D<E<L
20: course outline	.09	.09	.05	D<E=L
21: enjoyment	.14	.12	.19	E<L<D

[a]Order of numeric difference, not statistically significant difference.

[b]I = instructor.

[c]S = student.

that the least lenient response format would tend to have the greatest rater reliability. The inspection only succeeded in confirming the lack of significant differences among the three formats. The reliabilities of six items were compatible with the order L<E<D (four items L<E<D and two items L<E=D), while the reliabilities of six other items were compatible with the opposite order D<E<L (three items D<E<L and three items D<E=L). This meant that for six items, the descriptive format was most reliable, while for six others, the Likert format was most reliable. The evaluative format, which should have been most reliable if the hypothesized relationship between leniency and rater reliability were to be demonstrated, was moderately reliable in all 12 items. Other orderings were E<L=D four items compatible, and D<L<E one item compatible. No item reliabilities were ordered according to the order of means established by tests of leniency, L<D<E, though the two items with the order L<E=D could have been compatible with such an order. The lack of a trend in these results indicated that the conclusion of no differences in rater reliability between the three formats was completely accurate.

The ranges of single rater item reliabilities for each item type were quite similar to each other with the Likert format having a slightly greater range than the other two. The ranges were .01 to .34 for Likert format items,

.03 to .27 for descriptive format items, and .05 to .30 for evaluative format items. The median reliability values for the three formats were .11 for Likert, .13 for descriptive, and .13 for evaluative.

The graph of item reliabilities (Figure 4.2) illustrates the lack of superiority of any one item format over another in producing rater reliability. The amount of reliability appeared to be influenced more by the question asked than by the response cue format. Different factors appeared to produce high and low reliabilities across all formats. The factor of Student-Instructor Interaction was rated with the greatest reliability in all formats by the students, while the students in general were not as well able to rate their own interest or the course organization in any format. The similarity in performance among the three response types within the factors suggested that the factor, not the format, was the cause of the resulting reliabilities. No one item type was consistently superior within or across factors. The most uniformity occurred in the factor of Instructor Involvement where three of the four items were most reliably rated by the descriptive format items, the next most reliably rated by the evaluative format, and the least reliably rated by the Likert format.

The small size of the reliability estimates was due to the fact that they were single rater estimates.

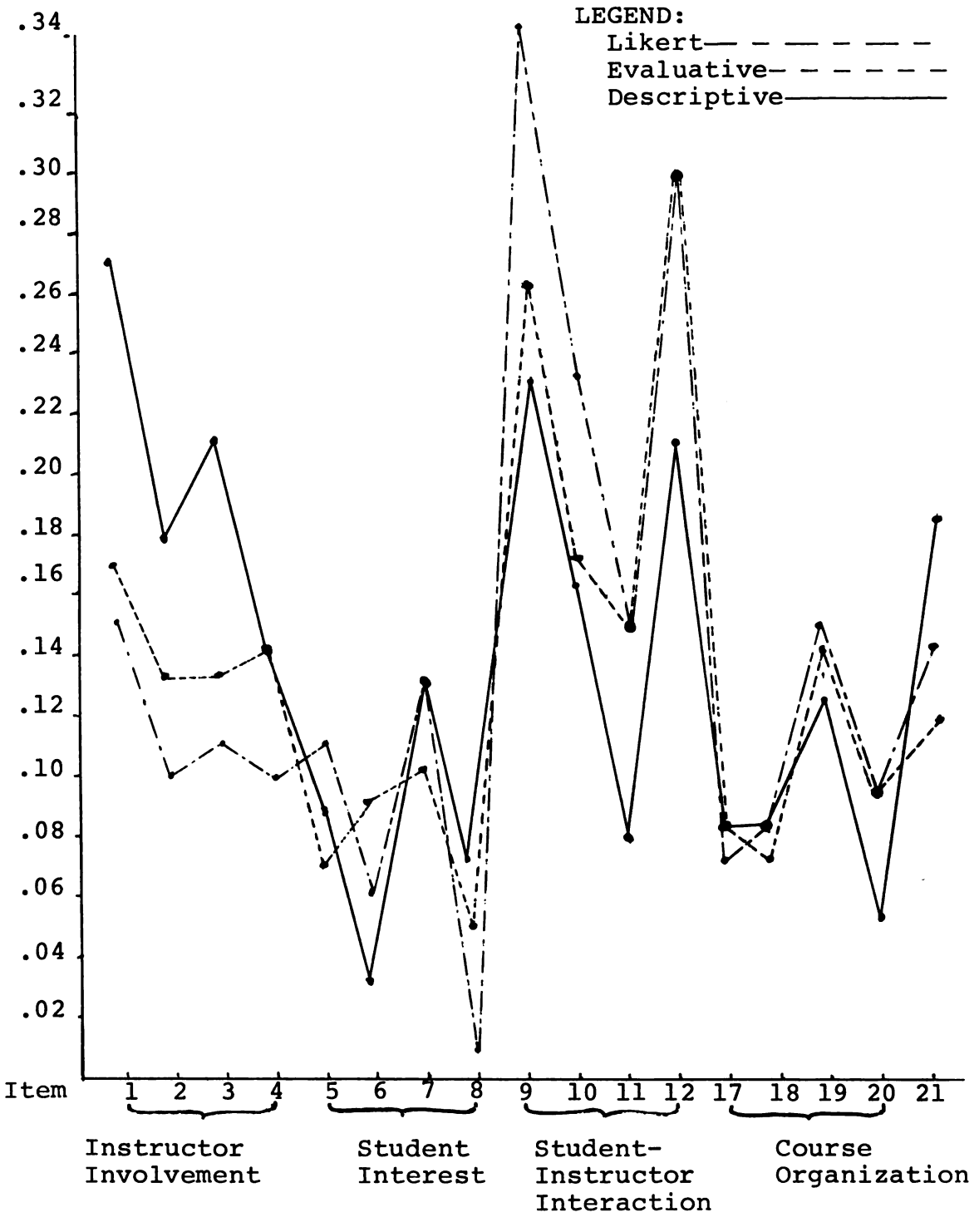

Figure 4.2.--Graph of item reliabilities for a single average rater.

Though these single rater estimates were quite valid for testing the significance of differences and for making comparisons of ranges and medians of the item types, the reliability of an average of 20 raters was calculated to determine the reliability of the instructional rating form items in use (Table 4.7). The Spearman-Brown formula was used to step up the single rater estimates. Remmers (1927, 1931) showed this to be a valid estimation technique

TABLE 4.7.--Item reliabilities for 20 raters.

	Item	Likert	Evaluative	Descriptive	Order[a]
Factor 1: Instructor Involvement					
1:	I-enthusiasm[b]	.79	.81	.88	L<E<D
2:	I-interest	.69	.74	.81	L<E<D
3:	I-examples	.70	.75	.84	L<E<D
4:	I-concern	.69	.77	.76	L<D<E
Factor 2: Student Interest					
5:	S-interest[c]	.72	.60	.67	E<D<L
6:	S-attention	.56	.85	.41	D<L<E
7:	S-challenge	.74	.69	.75	E<L<D
8:	S-competence	.20	.53	.61	L<E<D
Factor 3: Student Instructor Interaction					
9:	opinions	.91	.88	.86	D<E<L
10:	new ideas	.86	.80	.79	D<E<L
11:	questions	.78	.78	.64	D<E=L
12:	discussion	.90	.90	.84	D<E=L
Factor 4: Course Organization					
17:	unity of topics	.59	.64	.64	L<D=E
18:	organization	.64	.60	.62	E<D<L
19:	note-taking	.79	.77	.73	D<E<L
20:	course outline	.66	.66	.51	D<E=L
21:	enjoyment	.76	.72	.82	E<L<D

[a]Order of numeric difference, not statistically significant difference.

[b]I = instructor.

[c]S = student.

for student rating of instruction. The same order of results (except for rounding errors) is maintained among the stepped-up reliabilities as among the single rater estimates, but the stepped up values are more comparable to other classroom size rater reliability estimates.

Inspection of the rater reliabilities for 20 raters showed that the median item reliability for each of the three formats in a classroom situation was .72 for the Likert format and .75 for the descriptive and evaluative formats. These median values were slightly less than those reported by other university rating forms, probably because of the relatively small number of instructors taking part in the study. None of the reported values was based on a sample of less than 50 instructors, while this study was based on a sample of 23.

In summary, it was found that there were no differences in rater reliability between items with Likert, evaluative, and descriptive response cue formats in student rating of instruction. No trends were found among the data to support the hope that the least lenient response format would tend to have the greatest rater reliability. Rater reliabilities for all formats appeared to be consistently high or low according to the item being rated rather than the particular response format. The factor of Student-Instructor Interaction was found to be most reliably rated in all formats, while the factors

of Student Interest and Course Organization were not as reliably rated in any format. The item reliabilities for a single rater ranged from .01 to .34 for all formats with median reliabilities of .11 for Likert and .13 for descriptive and evaluative formats. When item reliabilities were stepped up by the Spearman Brown formula to show their likely performance in a classroom-size group, the median item reliabilities became .72 for the Likert format and .75 for the descriptive and evaluative formats.

Summary of Results of the Study

The hypothesis of no differences in mean ratings of instructors between items with Likert, evaluative, and descriptive response cue formats was rejected, establishing that there were significant differences in leniency between the three instructional rating forms. Inspection of the individual item F's indicated that the differences were likely present in most of the items. Scheffe post hoc analysis of the directions of the differences between individual item means indicated that the predicted directions of differences were partially correct. It was predicted according to claims made in the literature that the descriptive format would be the least lenient, the Likert format more lenient, and the evaluative format the most lenient. A contrast of Likert and descriptive item formats showed that the descriptive

format was in fact less lenient than the Likert format as predicted. But the ordering of the item means in the study data differed from the ordering predicted in that the evaluative format was found to be the least lenient of all formats instead of the most lenient. The evaluative format items had less lenient means than either the descriptive or Likert formats for the majority of items. The Likert format, which was the format of the rating scale currently in use at the university, was found to be the most often prone to leniency bias.

Inspection of the range of instructor means produced by each format corroborated the finding concerning leniency of the various formats. The evaluative format produced the smallest percentage of extremely lenient means while the Likert format produced the greatest percentage of them. It was noted that while there were statistically significant differences in lenient responding between the formats, the majority of instructor means in all formats was concentrated in the upper half of the scale. This was thought to be due in part to the higher quality of instructor who would volunteer to be evaluated.

Inspection of the graph of item means and of the questions themselves indicated that the topic of questions (the particular factor) was sometimes related to the type of response format that worked best (least

leniently) with it. Although the majority of the questions on all topics were answered least leniently in evaluative format, it was most successful with the topics of Course Organization and Student-Instructor Interaction. The descriptive format showed the most promise with questions concerning Instructor Involvement and Student Interest, equalling or surpassing the evaluative format means in half of the items. These results seemed to indicate that while evaluative items were the best overall rating measure, descriptive items were as good for rating people.

The claim concerning the superiority of multiple choice over fixed alternative items in reducing bias was not substantiated by this study of leniency. Fixed alternative Likert and evaluative response formats were found to produce both the most and least lenient responses respectively, while the multiple choice descriptive format was moderately successful in reducing lenient responding.

The hypothesis of no differences in rater reliability between items with Likert, evaluative, and descriptive response cue formats was not rejected, indicating that differences in the reliability of student ratings among the three formats were not large enough to rule out the possibility of their being due to chance. No trends were found among the data to support the hope that the least lenient response format would

tend to have the greatest rater reliability. Rater reliabilities for all formats appeared to be consistently high or low according to the item being rated rather than the particular response format. The factor of Student-Instructor Interaction was found to be most reliably rated in all formats, while the factors of Student Interest and Course Organization were not as reliably rated in any format. The item reliabilities for a single rater ranged from .01 to .34 for all formats with median reliabilities of .11 for Likert and .13 for descriptive and evaluative formats. When item reliabilities were stepped up by the Spearman-Brown formula to show their likely performance in a classroom-size group, the median item reliabilities became .72 for the Likert format and .75 for the descriptive and evaluative formats.

In Chapter V, the results are discussed in the light of the original purpose of the study and the claims made in the literature. Explanations are offered for the results observed and suggestions are made for the use of the three response formats in the future.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

Over the years many efforts have been made to improve student ratings of teacher effectiveness. This study represents another such effort. It is concerned with the particular problem of the leniency bias shown by many students in rating their instructors. By leniency bias is meant the tendency of students to use only the two or three highest options in rating their instructors. The harmful effect of this bias is to reduce discrimination between instructors to the extent that small differences in mean ratings produce large differences in reported rankings. The idea which gave rise to the present study was that leniency bias could be reduced by changing the wording of the response options. It was hoped that a different wording would increase the range of options used by student raters and improve discrimination between instructors. The major reason for conducting the study was to improve an existing Likert-type student instructional rating scale. Since the content of the scale was well established in its creation, the study was focussed on manipulating the response options

to reduce the amount of lenient responding present with the existing scale. Two alternative response definitions were chosen to compare with the existing Likert format response definitions. The descriptive graphic format was chosen as the most often recommended format for reducing leniency bias. It was hypothesized that this format would produce the least lenient responses from student raters of instruction. The evaluative format was chosen as a second alternative for purposes of contrast because it was claimed to be the most bias-prone response format. It was hypothesized that the evaluative format would produce the most lenient responses, and the descriptive the least lenient responses. It was also hypothesized that the least lenient response cue format would prove to have the greatest rater reliability.

A concurrent purpose of the study was to test two claims made in the literature concerning the bias-proneness of certain response definitions. The testing of the claims as they applied to the bias of lenient responding was compatible with the major objective of the study and so was included as part of the study. The claims tested were:

1. Evaluative cues are more susceptible to response bias than other cues.

2. Fixed response alternatives are more susceptible to bias than descriptive multiple choice alternatives.

If both of these claims were to hold true for leniency bias, the hypothesized order of response cue types, from most to least lenient would be, fixed alternative evaluative cues, fixed alternative Likert cues, and multiple choice short descriptive cues.

To conduct the study, three instructional rating forms differing primarily in response cue format were developed and administered to random thirds of the classes of 23 instructors. Leniency bias was measured by finding the closeness of each item mean to the midpoint of the rating scale. Since student ratings were overwhelmingly concentrated at the upper end of the scale, the format that gave the lowest mean was regarded as the least biased. The hypothesis of no differences in mean ratings (leniency bias) was tested with a two-way multivariate analysis of variance design, instructor by treatment, where the response cue formats were the three treatments and the 17 usable items were 17 dependent variables. Scheffe post hoc analyses tested alternate hypotheses that the evaluative format would produce the most lenient items, the Likert format the next most lenient, and the descriptive format the least lenient.

The hypothesis of no differences in rater reliabilities was tested by comparing confidence intervals about the reliability estimate for each item. Non-overlapping confidence intervals would indicate significant differences in rater reliabilities.

The hypothesis of no differences in mean ratings of instructors between items with Likert, evaluative, and descriptive response cue formats was rejected, establishing that there were significant differences in leniency between the three instructional rating forms. Inspection of the individual item F's indicated that the differences were likely present in most of the items. Scheffe post hoc analysis of the directions of the differences between individual item means indicated that the predicted directions of differences were partially correct. It was predicted, according to claims made in the literature, that the descriptive format would be the least lenient, the Likert format more lenient, and the evaluative format the most lenient. A contrast of Likert and descriptive item formats showed that the descriptive format was in fact less lenient than the Likert format as predicted. But the ordering of the item means in the study data differed from the ordering predicted in that the evaluative format was found to be the least lenient of all formats instead of the most lenient. The evaluative format items had less lenient means than either the descriptive or Likert formats for the majority of items. The Likert format, which was the format of the rating scale currently in use at the university, was found to be the most often prone to leniency bias.

Inspection of the range of instructor means produced by each format corroborated the finding concerning leniency-proneness of the various formats. The evaluative format produced the smallest percentage of extremely lenient means while the Likert format produced the greatest percentage of them. It was noted that while there were statistically significant differences in lenient responding between the formats, the majority of instructor means in all formats was concentrated in the upper half of the scale. This was thought to be due in part to the higher quality of instructor who would volunteer to be evaluated.

Inspection of the graph of item means and of the questions themselves indicated that the topic of questions (the particular factor) was sometimes related to the type of response format that worked best (least leniently) with it. Although the majority of the questions on all topics were answered least leniently in evaluative format, it was most successful with the topics of Course Organization and Student-Instructor Interaction. The descriptive format showed the most promise with questions concerning Instructor Involvement and Student Interest, equalling or surpassing the evaluative format means in half of the items. These results seemed to indicate that while evaluative items were the best overall rating measure, descriptive items were as good for rating people.

The claim concerning the superiority of multiple choice over fixed alternative items in reducing bias was not substantiated by this study of leniency. Fixed alternative evaluative and Likert response formats were found to produce both the most and least lenient responses in the study, while the multiple choice descriptive format was moderately successful in reducing lenient responding.

The hypothesis of no differences in rater reliability between items with Likert, evaluative, and descriptive response cue formats was not rejected, indicating that differences in the reliability of student ratings among the three formats were not large enough to rule out the possibility of their being due to chance. No trends were found among the data to support the hope that the least lenient response format would tend to have the greatest rater reliability. Rater reliabilities for all formats appeared to be consistently high or low according to the item being rated rather than the particular response format. The factor of Student-Instructor Interaction was found to be most reliably rated in all formats, while the factors of Student Interest and Course Organization were not as reliably rated in any format. The item reliabilities for a single rater ranged from .01 to .34 for all formats with median reliabilities of .11 for Likert and .13 for descriptive and evaluative formats. When item reliabilities were

stepped up by the Spearman-Brown formula to show their likely performance in a classroom-size group, the median item reliabilities became .72 for the Likert format and .75 for the descriptive and evaluative formats.

Conclusions

1. Evaluative format items in instructional rating scales were less prone to leniency bias and had rater reliabilities comparable to Likert and descriptive formats, making them the best choice of the three formats to improve the existing instructional rating form.

2. Claims made in the literature concerning the proneness to bias of fixed alternative response formats in general, and evaluative formats in particular, were found not to hold with student ratings of instruction.

   a. The evaluative format was not the most bias-prone response format as claimed in the literature. This study of leniency bias found it to be the least lenient of three response formats--Likert, descriptive, and evaluative.

   b. Fixed response alternative were not more susceptible to leniency bias than descriptive multiple choice alternatives for student ratings of instruction. Fixed alternative evaluative response cues were found to be least susceptible to leniency bias in this study, while multiple choice descriptive response cues were found to

be moderately susceptible to leniency bias, and fixed alternative Likert items most susceptible.

3. The reduction in lenient responding produced by the evaluative format items was not large enough to increase the range of scale values used by raters and improve discrimination between instructors to the extent that a noticeable increase in rater reliability would result.

4. The rater reliabilities for the three forms were within an acceptable range for use with classroom-size groups of raters.

5. The topic of questions (the particular factor) was sometimes related to the type of response format that worked best (least leniently) with it. Though the evaluative format was the least lenient format for the majority of items, it was found to be most successful with the topics of Course Organization and Student-Instructor Interaction, while the descriptive format was found to be as good as the evaluative format for half the items concerning Instructor Involvement and Student Interest.

Discussion

The study showed that leniency bias in the existing Likert-type instructional rating scale could be reduced by using a different response cue format. It was contrary to expectations, however, that the evaluative format should be found to be the best choice. In

searching for an explanation of this phenomenon it became apparent that several variables in the student rating of instruction situation made it different from the traditional rating situations dealt with by measurement specialists in the past. First, instead of a small number of trained raters as in the typical rating situation, there was a large number of untrained raters. The large numbers would tend to insure the reliability of the results, but the fact that they were untrained would influence their ability to rate accurately with a particular response format. The response format with which students were most familiar from years of past experience with grading was the evaluative format. Since the raters were untrained, their common past experiences as students became an influential variable in the results of this study. Secondly, the purpose of the scale as an institutional instrument to evaluate instruction and the non-specific nature of the questions which had to be general enough to apply to all types of classrooms lent itself more to the evaluative response mode than might a scale created by the instructor to diagnose the difficulties in his particular class. In other situations, where specific behavioral questions were asked rather than general summarizing questions such as those in the form used in this study, a different format, such as the descriptive format, might be more appropriate. But for

this rating task, it appeared that the evaluative format was the best fit to the purpose of the scale. Third, the conditions under which the scale is administered to obtain the student ratings are often rushed or pressured, so that a simple, easily digested wording of responses common to all questions would be more easily used by the student raters than one in which they would be required to digest the meaning of a new response continuum on each question. For trained or unpressured raters the variable wording of response options of the descriptive format might be beneficial since it would reduce boredom and the chance of getting in a rut answering several questions. But in the instructional rating situation where speeded responses may be a greater factor than the possibility of boredom, the more easily digested fixed alternative format appeared to have the advantage. Of course, this line of reasoning does not explain why the fixed alternative Likert format was not more successful than it was, since in general it was the most prone to leniency bias of the three formats, but the original variable of students' greater experience with evaluative than Likert formats helps account for this. A further explanation for the poor showing of the Likert format items was found in the literature concerning the ambiguity of the neutral response option in rating scales in general. Holdaway's result showing that about as many

people chose the neutral option as chose the disagree option when the neutral was not there indicates that a response of neutral probably has more negative than neutral connotations, limiting the number of favorable-sounding choices in the Likert scale with a neutral midpoint to the top two. This would serve to exaggerate any tendency to leniency in ratings that might already be there, causing the poor showing of the Likert format in instructional rating form items.

Although significant differences in lenient responding were found among the three response formats, the study was unable to show corresponding differences in rater reliability. It was expected that a reduction in lenient responding would increase the range of scale values used by raters and thus increase the differences in mean ratings between good and poor instructors. A look at the range of instructor means for the three response formats indicated that the evaluative format did in fact produce a larger number of less-than-midpoint instructor means than either the Likert or descriptive formats, but that the majority of means in all formats was still within the upper half of the scale. Thus, although there were significant differences in lenient responding, the magnitude of the differences was not sufficient to significantly increase the range of instructor means and improve the resulting rater reliability

estimates. The lack of significant differences in rater reliability in this study was not assumed to contradict the rationale behind reducing lenient responding to improve rater reliability, but it did indicate that the size of the reduction actually obtained in this study was not large enough to create the desired effect.

The median rater reliabilities obtained for all of the response formats in the study were thought to be comparable to those reported by other collegiate instructor scales. The variables influencing these estimates were the number of student raters assumed in the calculation and the number of instructors contributing ratings to the study. Most collegiate scales reported rater reliabilities based on 20 or more students per class and 50 or more instructors. Rater reliabilities obtained in this fashion ranged from .87 based on 205 instructors to .79 based on 32 instructors, all with 20 or more students per class. All reported reliabilities greater than .80 were based on 50 or more instructors. It appeared that the added variance created by a larger sample size improved the reliability estimate obtained. Since the reliability estimates of this study were based on 23 instructors, and volunteers at that, the median values for 20 raters of .72 and .75 were seen to be within an acceptable range for use with classroom-size groups of raters. There

was reason to believe that the estimates would have been improved if based on a larger number of instructors.

Possible differences in performance of the response formats between the different topics of the rating scale questions were not statistically tested in this study, but trends were noted which suggested that the relationship between the topic of the question and response format performance with it was worthy of further exploration. Though the evaluative format was the least lenient format for the majority of items in this instructional rating scale, it was found to be the most successful with the topics of Course Organization and Student-Instructor Interaction, while the descriptive format was found to be as good as the evaluative format for half the items concerning Instructor Involvement and Student Interest. The relationship was too tenuous for a conclusive statement to be based upon it, but it might be proposed that the closer the match of topic and response mode, the less prone were the ratings to bias and unreliability. As a hypothesis in this direction, it might be expected that descriptive items would work best in the rating of individual instructor and student behaviors to be used as part of the diagnosis of instructional problems, while the evaluative format would work best in the rating of general aspects of the course for purposes of departmental accountability. Since the existing rating

form was attempting to serve both purposes, it was not surprising from this point of view that the evaluative format worked best for most items, with some comparable performances by the descriptive format in the areas of Instructor Involvement and Student Interest.

In all, it was found that the study was partially successful in obtaining its ends--succeeding in reducing lenient responding by changing the response mode, but failing to reduce it sufficiently to improve the rater reliability of the instructional rating form. The claims made in the literature concerning fixed response alternatives in general, and the evaluative format in particular, were found not to hold in the student rating of instruction situation. The evaluative format items of the instructional rating scale were found to be least prone to leniency bias, comparable in rater reliabilities to Likert and descriptive formats, and most consistent with the experiences of the raters and the normative purposes of the rating task. It was concluded therefore that they were the best choice of the three formats to improve the existing instructional rating scale.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aleamoni, L. M. "The Evaluation of Instruction." Draft, November 15, 1972.

Bendig, A. W. "Ability and Personality Characteristics of Introductory Psychology Instructors Rated Competent and Empathetic by Their Students." Journal of Educational Research, XLVIII (1955), 705-9.

Borg, W. R. "Personality and Interest Measures as Related to Criteria of Instructor Effectiveness." Journal of Educational Research, L (1957), 701-9.

Brogden, H. E., and Taylor, E. K. "The Theory and Classification of Criterion Bias." Educational and Psychological Measurement, X (1950), 159-86.

Brown, B. B.; Mendenhall, W.; and Beaver, R. "The Reliability of Observations of Teachers' Classroom Behavior." Journal of Experimental Education, XXXVI (1968), 1-8.

Bryan, R. C. "Comparison of Two Instruments for Use in Evaluating Pupil Reactions." School Review, LII (1944), 285-92.

Caldwell, E. "A Review of Student Rating of Instruction at U.S.F. and Elsewhere." University of South Florida Institutional Report No. 59, Office of Evaluation Services, March 12, 1971.

Centra, John A. "The Student Instructional Report: Its Development and Uses. (SIR Report No. 1)." Princeton, New Jersey: Educational Testing Service, 1972.

Champney, H. "The Measurement of Parent Behavior." Child Development, XII (1941), 131-66.

Coffman, W. E. "Determining Students' Concepts of Effective Teaching From Their Ratings of Instructors." Journal of Educational Psychology, XLV (1954), 277-86.

Costin, F. "A Graduate Course in the Teaching of Psychology: Description and Evaluation." Journal of Teacher Education, XIX (1968), 425-32.

_____; Greenough, W. T.; and Menges, R. J. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness." Review of Educational Research, XLI (1971), 511-35.

Cronbach, L. J. "Response Sets and Test Validity." Educational and Psychological Measurement, VI (1946), 475-94.

_____. "Further Evidence on Response Sets and Test Design." Educational and Psychological Measurement, X (1950), 3-31.

_____. Essentials of Psychological Testing. 2nd ed. New York: Harper, 1960.

_____; Rajaratnam, N.; and Gleser, G. C. "Theory of Generalizability: a Liberalization of Reliability Theory." British Journal of Statistical Psychology, XVI (1963), 137-63.

Deshpande, A. S.; Webb, S. C.; and Marks, E. "Student Perceptions of Engineering Instructor Behaviors and Their Relationships to the Evaluation of Instructors and Courses." American Educational Research Journal, VII (1970), 289-305.

Ebel, R. L. "Estimation of the Reliability of Ratings." Psychometrika, XVI (1951), 407-24.

Edwards, A. L. The Measurement of Personality Traits by Scales and Inventories. New York: Holt, Rinehart and Winston, Inc., 1970.

Elliott, L. L. "Effects of Item Construction and Respondent Aptitude on Response Acquiescence." Educational and Psychological Measurement, XXI (1961), 405-15.

Engelhart, M. D. "A Method of Estimating the Reliability of Ratings Compared with Certain Methods of Estimating the Reliability of Tests." Educational and Psychological Measurement, XIX (1959), 579-88.

Finn, J. D. "Univariate and Multivariate Analysis of Variance and Covariance: A Fortran IV Program." Modified for the Michigan State University CDC 3600 and 6500 computer systems by David J. Wright, Office of Research Consultation, March, 1970.

Finn, R. H. "A Note on Estimating the Reliability of Categorical Data." Educational and Psychological Measurement, XXX (1970), 71-6.

_____. "Effects of Some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings." Educational and Psychological Measurement, XXXII (1972), 255-65.

Follman, J. "NOTUP (New Observation of Teaching of University Professors) Student Rating Scale Key." Submitted to Phi Delta Kappan.

Gillmore, G. M. "Three Functions of Student Course Evaluations." Paper presented at a Symposium on Course Evaluation, NCME convention, 1972.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill Book Company, 1954.

_____, and Jorgensen, A. P. "Some Constant Errors in Ratings." Journal of Experimental Psychology, XXII (1938), 43-57.

Hildebrand, M.; Wilson, R. C.; and Dienst, E. R. "Evaluating University Teaching." Center for Research and Development in Higher Education, University of California, 1971.

Hodgson, T. F. "The General and Primary Factors in Student Evaluation of Teaching Ability." Seattle, Washington: University of Washington, 1958.

Holdaway, E. A. "Different Response Categories and Questionnaire Response Patterns." Journal of Experimental Education, XL (1971), 57-60.

Levinthal, C. F.; Lansky, L. M.; and Andrews, C. E. "Student Evaluations of Teacher Behaviors as Estimations of Real-Ideal Discrepancies: A Critique of Teacher Rating Methods." Journal of Educational Psychology, LXII (1971), 104-9.

Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton Mifflin, 1953, Chapter 16.

Lovell, G. D., and Haner, C. F. "Forced-Choice Applied to College Faculty Rating." Educational and Psychological Measurement, XV (1955), 291-304.

Madden, J. M., and Bourdon, R. D. "Effects of Variations in Rating Scale Format on Judgment." Journal of Applied Psychology, XLVIII (1964), 147-51.

Matell, M. S., and Jacoby, J. "Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity." Educational and Psychological Measurement, XXXI (1971), 657-74.

Medley, D. M., and Mitzel, H. E. "Measuring Classroom Behavior by Systematic Observation." Handbook of Research on Teaching. Edited by N. L. Gage. Chicago: Rand-McNally and Company, 1963, Chapter 6.

Office of Evaluation Services, MSU. "Student Instructional Rating System (SIRS) Technical Bulletin." December 22, 1969.

_____. "Student Instructional Rating System Analysis of Responses for Winter Term 1970." SIRS Research Report #1, February 15, 1971.

Oppenheim, A. N. Questionnaire Design and Attitude Measurement. New York: Basic Books, Inc., 1966.

Peters, D. L., and McCormick, E. J. "Comparative Reliability of Numerically Anchored Versus Job-Task Anchored Rating Scales." Journal of Applied Psychology, L (1966), 92-6.

Remmers, H. H. "The Equivalence of Judgments and Test Items in the Sense of the Spearman-Brown Formula." Journal of Educational Psychology, XXII (1931), 66-71.

_____. "Reliability and Halo Effect of High School and College Students' Judgments of Their Teachers." Journal of Applied Psychology, XVIII (1934), 619-30.

_____; Shock, N. W., and Kelly, E. L. "An Empirical Study of the Validity of the Spearman-Brown Formula as Applied to the Purdue Rating Scale." Journal of Educational Psychology, XVIII (1927), 187-95.

_____, and Weisbrodt, J. A. Manual of Instructions for the Purdue Rating Scale for Instruction. Revised ed. 1965. West Lafayette, Indiana: University Book Store, 1965.

Rezler, A. G. "The Influence of Needs Upon the Student's Perception of His Instructor." Journal of Educational Research, LVIII (1965), 282-6.

Ryans, D. G. Characteristics of Teachers: Their Description, Comparison, and Appraisal. Washington, D. C.: American Council on Education, 1960.

Sakoda, J. M.; Cohen, B. H.; and Beall, G. "Test of Significance For a Series of Statistical Tests." Psychological Bulletin, LI (1954), 172-5.

Sharon, Amiel. "Eliminating Bias from Student Ratings of College Instructors." Journal of Applied Psychology, LIV (1970), 278-81.

Smith, D. H. "Correcting for Social Desirability Response Sets in Opinion-Attitude Survey Research." Public Opinion Quarterly, XXXI (1967), 87-94.

Spencer, R. E., and Aleamoni, L. "The Illinois Course Evaluation Questionnaire: A Description of Its Development and a Report of Some of Its Results. Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1969.

Stanley, J. C. "Reliability." Educational Measurement. 2nd ed. Edited by R. L. Thorndike. Washington, D.C.: American Council on Education, 1971, Chapter 13.

Stockford, L., and Bissell, H. W. "Factors Involved in Establishing a Merit-Rating Scale." Personnel, XXVI (1949), 94-118.

Stuit, D. B., and Ebel, R. L. "Instructor Rating at a Large State University." College and University, XXVII (1952), 247-54.

Swanson, R. A., and Sisson, D. J. "The Development, Evaluation, and Utilization of a Departmental Faculty Appraisal System." Journal of Industrial Teacher Education, IX (1971), 64-79.

Symonds, P. M. Diagnosing Personality and Conduct. New York: Appleton-Century-Crofts, 1931.

Taylor, J. B.; Ptacek, M.; Carithers, M.; Griffin, C.; and Coyne, L. "Rating Scales as Measures of Clinical Judgment III: Judgments of the Self on

Personality Inventory Scales and Direct Ratings." Educational and Psychological Measurement, XXXII (1972), 543-57.

Thorndike, R. L., and Hagen, E. P. Measurement and Evaluation in Psychology and Education. 2nd ed. New York: John Wiley and Sons, Inc., 1961.

Tyler, T. A. "Reducing the Threat to Instructors in Evaluation of Instruction Programs." Paper presented at a Symposium on Course Evaluation, NCME convention, 1972.

APPENDIX

APPENDIX

RATER RELIABILITY CONFIDENCE INTERVALS FOR r_{11}

$\alpha = .05$

Item / Form	1	2	3	4
LIKERT	.075 - .258	.037 - .185	.041 - .194	.038 - .188
EVAL.	.089 - .282	.056 - .222	.059 - .228	.067 - .242
DESC.	.161 - .397	.094 - .292	.116 - .328	.065 - .240
	5	6	7	8
LIKERT	.047 - .205	.010 - .130	.057 - .224	-.020 - .061
EVAL.	.018 - .146	.031 - .172	.039 - .189	.007 - .122
DESC.	.032 - .176	-.006 - .094	.058 - .227	.017 - .145
	9	10	11	12
LIKERT	.215 - .470	.134 - .356	.071 - .250	.188 - .433
EVAL.	.156 - .389	.084 - .273	.072 - .251	.187 - .432
DESC.	.132 - .353	.078 - .264	.024 - .161	.114 - .324
	17	18	19	20
LIKERT	.015 - .141	.026 - .164	.076 - .259	.029 - .170
EVAL.	.025 - .160	.018 - .145	.069 - .246	.030 - .171
DESC.	.025 - .162	.021 - .153	.050 - .211	.004 - .117
	21			
LIKERT	.062 - .235			
EVAL.	.049 - .208			
DESC.	.101 - .302			