30 (

RATING SCALE FORMAT AND THE EFFECTIVENESS OF TRAINING RATERS TO MINIMIZE RATING ERRORS

by

LOREN K. KIRKEIDE

B. A., Minot State College, 1974

A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Psychology

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1980

Approved by:

Major Professor

ACKNOWLEDGMENTS

The author wishes to express his appreciation to Professor Arthur Dayton for his assistance in the statistical analysis of the data. Appreciation is also due to the many friends who have offered their help and support. Finally, I wish to acknowledge the invaluable help and support of my wife.

TABLE OF CONTENTS

LIST OF TABLES

Introduction

General Statement

Performance evaluations have played a very pervasive role in organizations as feedback for improving employee job performance and as criteria in the areas of selection, training, wage and salary administration, transfer, promotion, demotion, and dismissal. With this wide range of application and substantial contribution to organizational effectiveness, grave concern over the validity of performance evaluations is unquestionably warranted, as inclusion of inadequate and/or invalid performance measures would surely compromise the value of any outcomes emerging from procedures wherein the ratings serve an integral function. Consequently, this necessitates an extensive, systematic approach to the development of comprehensive, reliable, valid and, of course, acceptable performance criteria, as opposed to criteria established essentially by fiat.

Performance in many jobs and activities surfaces as a highly complex and seemingly intangible construct, defying any simplistic level of operationalization. Its multidimensional nature usually precludes the use or limits the value of any single overall measure of performance, and purely objective performance measures do not seem to effectively capture the qualitative aspects inherent in certain job requirements. Thus, organizations have tended to rely more upon the subjective judgments of individuals concerned with and responsible for the performance of others as a more acceptable assessment (Guion, 1965; Landy and Trumbo, 1976).

The elicitation of subjective judgments through numerical ratings on a rating scale is by no means new. According to Ellson and Ellson (1953), the rating scale was in use as early as 1826 to evaluate such characteristics as judgment, imagination, perception, and courage. While such ratings might appear to increase the validity of performance evaluations, their accuracy in reflecting "true" performance is also far from incontestable. Use of subjective ratings to resolve the problem of criterion deficiency in objective performance measures maximizes the likelihood of criterion contamination, resulting from inconsistent operational definitions among raters for considerations that are extremely difficult to measure (e.g., personality traits), and from the inclusion of rater considerations that are irrelevant to job performance. Thus, to alleviate notoriously inaccurate evaluations, raters are persuaded to base their subjective ratings on objective performance standards (Borman, 1975; Bernardin and Walter, 1977; Bernardin, 1978).

Attempts to minimize irrelevant considerations as well as improve the assessment of relevant considerations have led to the development of several innovative techniques for evaluating performance (e.g., behavioral expectation scales, mixed standard scales). The effectiveness of each technique is generally gauged by its ability to improve the psychometric quality of the performance ratings. Since it is impossible to actually distinguish between erroneous and accurate performance evaluations, a rating scale's superiority has therefore been contingent on the extent to which the scale demonstrates comparatively better psychometric properties with respect to halo, leniency/severity, central tendency/range restriction, and inter-rater agreement.

Halo error can be defined conceptually as a rater tendency to evaluate the performance of an individual in several ostensibly independent performance areas in a manner that is consistent with her/his general, overall impression of the ratee, rather than evaluate ratee performance on each dimension only according to performance concerning the dimension (Saal, Downey, and Lahey, Note 2).

Leniency/severity can be conceptualized as a "level" effect. This error refers to a rater tendency to give either generally high ratings to all ratees or generally low ratings to all ratees, regardless of the actual performance exhibited by each ratee (Saal et al., Note 2).

Central tendency and restriction of range refer to a rater tendency to give similar ratings to all ratees. That is, the ratings do not effectively differentiate the ratees with respect to relative levels of performance.

While restriction of range can apply to any part of the scale, Saal et al. (Note 2) suggested that, in order to avoid potential confusion, the term "central tendency" should be restricted to instances where restriction of range occurs at the midpoint of the rating scale.

Interrater agreement refers to the extent to which raters independently provide similar ratings of the same ratee (Saal et al., Note 2).

There is clearly some confusion, however, concerning the operational as well as the conceptual definitions of the psychometric constructs (Downey and Saal, Note 1). Saal et al. (Note 2) have further demonstrated that the disparate operational definitions of a single construct can yield inconsistent interpretations of the same data. Consequently, until such time as uniform conceptual and operational definitions of these constructs become established, specification of the method of measurement is most important in any discussion referring to them.

Researchers have attempted to improve the psychometric quality of performance ratings by either developing different scale formats or training raters to rate more effectively. Two relatively new scale formats that have recently been given considerable attention in the literature are behavioral expectation scales (BES) and mixed standard scales (MSS). The respective potential advantages of each format in enhancing the accuracy and subsequent psychometric quality of the performance ratings are reviewed below, as are the potential drawbacks and available empirical research.

The efficacy of training raters to improve the quality of their ratings has also gained more attention in the literature. A discussion of the potential advantages and drawbacks of rater training is presented below, along with a review of related empirical research.

Behavioral Expectation Scales (BES)

Smith and Kendall (1963) introduced a procedure designed to develop job-related, behaviorally oriented appraisal standards directly related to performance standards required on the job. Included in the method is a variation of the critical incident technique proposed by Flanagan (1954). Critical incidents are defined as direct observations of human behavior that describe various levels of proficiency/effect in relation to the activity, and are critical to the outcome of the activity. Smith and Kendall (1963) used the incidents as bases for inferences/predictions of expected job performance. While the method for developing behavioral expectation scales (BES), also referred to as behaviorally anchored rating scales (BARS), has been employed with various procedural modifications, the design basically consists of five steps. A review of each step is reported as Appendix I. (The distinction between BES and BARS is also described in Appendix I.)

Potential Advantages of BES

A number of people have emphasized several advantages of this system over more conventional methods of scale development (Blood, 1974; Campbell, Dunnette, Arvey, and Hellervik, 1973; Campbell, Dunnette, Lawler, and Weick, 1970; Landy and Trumbo, 1976; Schwab, Heneman, and DeCotiis, 1975; Smith and Kendall, 1963). First, the conceptually independent dimensions and definitions are generated by a sample of the rater population, and not by an "outsider" attempting to derive them in a manner that minimizes direct rater input. The underlying rationale here is that raters are best qualified to determine aspects of the job to be evaluated, as well as those to be disregarded. There may be some question, however, as to whether raters are any more qualified than ratees (Saal, Note 4).

Second, the behavioral anchors offer raters who eventually use the scales something concrete with which to compare observed behaviors. This should help focus rater attention on each ratee's job-related behavior and away from any general, overall impression of the ratee.

Third, the anchors provide a mutual frame of reference for all raters, thereby allowing each rater to compare personal performance standards with consensual standards. Raters with personal standards that are incompatible with consensual standards can try to adopt the mutually accepted standards, if compelled to do so. Using only behavioral descriptions that satisfy both the retranslation and scaling criteria, however, insures a certain level of rater agreement concerning both the dimension that each anchor describes and the performance level. This should minimize ambiguity in the anchors and dimensions, and also minimize discrepancies among various raters' perceptions of the performance levels reflected by the anchors.

Fourth, continual use of the raters' terminology throughout the procedure should strengthen the construct validity of the emergent scales, as well as the "face" validity.

Together, these advantages might be expected to increase interrater agreement and minimize rating errors. Blood (1974) further suggested that the procedure may also be used to 1) generate behavioral statements for feedback to ratees, 2) extend the domain of evaluated performance, 3) aide in the development of training programs, 4) aide in assessing consensus on organizational policy, and 5) aide in assessing the accuracy of communication of organizational policy.

Potential Drawbacks of BES

First, selection criteria for behavioral statements in the retranslation and scaling phases might be too lenient to satisfactorily eliminate potentially ambiguous anchors, as current standards are set arbitrarily by each researcher. Consequently, a number of raters may strongly disagree with the anchors and choose to ignore them. Such disagreement would be minimized with sufficiently stringent standards.

Second, implicit in the use of BES is the assumption that raters are able to accurately infer expected behaviors from behaviors actually observed. Smith and Kendall (1963) believed that such an assumption was reasonable, provided one is dealing with a fairly homogeneous group of raters (i.e., individuals holding the same supervisory position). Whether or not such an assumption is indeed presumptuous, however, remains unknown. Some raters may be able to make such inferences, whereas others may not.

Third, the standard deviations associated with the scale values of many behavioral statements readily indicate substantial disagreement among raters regarding the level of performance exemplified by each statement. Given the large number of statements generated initially, only a small percentage survives the selection process. This is partly a function of the disagreement among raters concerning the performance level of each behavior. If, in fact, ratee behavior in certain jobs is replete with ambiguity, and consensual standards of performance are not easily detected, the value of behavioral anchors as reference points is diminished. Some raters may associate an ambiguous behavior with scaled expected behavior exemplifying moderate performance, whereas others may associate that behavior with scaled expected behavior exemplifying exceptionally good performance. Raters are therefore unable to reliably identify ratees' strong and weak performance areas with any acceptable level of consensus, as mutually agreed upon performance levels exhibited by a ratee are obscured by the far more prevalent ambiguous behaviors of the ratee. Lack of interrater agreement, for example, would not be due to poorly scaled expected behaviors, but rather to the predominance of ambiguity in ratee behavior, the desirability of which raters cannot collectively discern. If, on the other hand, ratee behavior is not exceedingly ambiguous, similar ratings are more likely to occur, provided that raters are able to infer expected behavior from observed behavior.

Research on BES

The inception of BES triggered a great deal of research on the procedure in several different areas. A series of studies was conducted by Maas (1965), in which interrater reliabilities of interviewers using conventional adjective rating scales for selecting orientation and dorm counselors were compared with interrater reliabilities of interviewers using BES. Interrater reliability (correlation of candidate ratings between two interviewers) was significantly better with BES.

Other studies also provided support for BES by reporting very high scale reliabilities (Fogli, Hulin, and Blood, 1971; Landy and Guion, 1970; Smith and Kendall, 1963). It was pointed out, however, that such reliabilities were spuriously high estimates, and that a scale's worth lies rather in its operational effectiveness (Borman and Vallon, 1974; Fogli, et al., 1971; Zedeck and Baker, 1972). Borman and Vallon (1974) described the computation of such scale reliabilities and listed techniques that artificially improved them.

Zedeck and Baker (1972) performed a multitrait-multirater analysis of nursing performance with BES (Campbell and Fiske, 1959; Lawler, 1967) and reported moderate convergent validity, but no discriminant validity. Since the raters represented two different and distinct supervisory levels, the authors suggested that each level possibly valued the various dimensions quite differently and also lacked equal opportunity to observe ratee behavior.

Campbell et al. (1973) responded to a need for comparative studies with BES by examining the relative effects of BES and summated rating scales (SRS) for department managers. SRS were constructed by breaking down the dimension definitions of BES "into their major elements, and each of these separate statements was used as a Likert-type item with a 4-point response format. All the items were statements or functions that contributed to high performance on a particular dimension, and the individual was rated as exhibiting it very rarely (1) to almost always (4)." (p. 18). BES was found superior in demonstrating less leniency error (mean ratings on each dimension), less halo error (clearer factor-analytic solution), and greater discriminant validity in a multitrait-multimethod analysis (Campbell and Fiske, 1959).

Subsequent comparative studies, however, have not been as conclusive regarding the superiority of BES. Borman and Vallon (1974) compared the original BES for nurses developed by Smith and Kendall (1963) with less complex graphic rating scales (GRS). The only difference between scales was that the behavioral anchors had been eliminated from GRS. Significantly less leniency error (mean dimension ratings) and significantly greater rater discrimination among ratee performance (standard deviation of each rater's ratings across ratees on a single dimension and administration) was found with the GRS. There were no significant differences with respect to halo error (dimension intercorrelations) and interrater agreement (correlation of ratings between head and assistant head nurses). In defense of BES, they suggested that since the dimensions and anchors were developed approximately ten years earlier, the rating scale may have simply been outdated.

Studies by Burnaska and Hollmann (1974) and Borman and Dunnette (1975) compared BES, GRS (BES without anchors), and a format in use prior to the time of their respective studies. The former study indicated no superiority of BES for rating college professors over the other two scales (analyses involved an ANOVA framework suggested by Guilford, 1954). The latter study reported BES for rating junior navy officers as somewhat superior to the others with respect to halo (variance in each rater's ratings across dimensions), leniency (mean dimension ratings), interrater agreement (correlation of ratings between commanding and executive navy officers), and rater differentiation among ratees' performance (standard deviation of each rater's ratings across ratees on a single dimension). The magnitudes of these differences, however, were quite small.

Keaveny and McGann (1975) reported that BES for rating college professors showed less halo error (dimension intercorrelations with raters as data points, dimension correlations with an overall performance rating) and greater interrater agreement (standard deviation of the ratings on each dimension for each ratee) than GRS (BES without anchors). They found no differences with respect to leniency error (mean dimension ratings).

A question arises concerning the appropriateness of their design for administering the scales, wherein all raters were given BES immediately following their ratings using GRS. In this design ratings on BES were vulnerable to bias from the preceding ratings on GRS, as the only difference between the scales was the presence or absence of anchors. Such a within-groups design is susceptible to confounded results if investigators are either not cognizant of or are careless in their efforts to minimize these problems. As Zedeck, Kafry, and Jacobs (1976) pointed out, "having raters use both formats may encourage an attempt by the rater to be consistent from format to format" (p. 173), especially when there is very little difference between the rating procedures.

Studies by Bernardin, Alvares, and Cranny (1976) and Bernardin (1977) recompared BES to summated rating scales (SRS) for rating college professors to determine the relative psychometric properties of each. The results revealed that BES and SRS performed equally well with respect to halo (first study calculated the dimension intercorrelations with raters as data points for each ratee, second study calculated the standard deviation of ratings across dimensions for each rater), leniency (mean dimension ratings), ratee discrimination (standard deviation of mean ratings by raters across ratees for each dimension), and interrater agreement when each scale underwent rigorous development, using optimal developmental and rating strategies. As indices of interrater agreement, both studies calculated for each ratee the standard deviation of ratings on each dimension. In addition, Bernardin et al. (1976) calculated for each ratee the correlation between raters with dimension ratings as data points, while Bernardin (1977) randomly selected a pair of raters from each of the ratees and, for each dimension, correlated the ratings of half of the raters (one from each pair) with the ratings of their respective partners. Procedures for developing BES were in accord with recommendations offered by Bernardin, LaShells, Smith, and Alvares (1976), while SRS were constructed from either the item-analyzed components of the performance dimension definitions (SRS_1) or the behavioral expectation items that had survived both the retranslation and scaling phases (SRS_2), as the retranslation phase is analogous to an item analysis.

DeCotiis (1977) also reported that BES for appraising vignettes describing patrol officer job performance performed no better psychometrically than two independently developed graphic and trait rating scales.

Based on the findings above, it appears that BES have not performed as well psychometrically as had been anticipated.

Several studies concerned specifically with various manipulations of the original Smith and Kendall (1963) system have investigated modifications in the developmental and scoring procedures to determine their effects regarding internal (e.g., Guttman scaling properties) as well as external psychometric considerations (Arvey and Hoyle, 1974; Bernardin, et al., 1976; Campion, Greener, and Wernli, 1973; Finley, Osburn, Dubin, and Jeanneret, 1977; Goodale and Burke, 1975; Kafry, Zedeck, and Jacobs, 1976; Zedeck, et al., 1976), and examined the psychometric properties of BES as a function of differential group characteristics (Cascio and Valenzi, 1977; Friedman and Cornelius, 1976; Schneier, 1977). Some have found the system useful as a means of detecting and assessing rater differences related to organizational level (Borman, 1974; Klimoski and London, 1974; Zedeck, Imparato, Krausz, and

Oleno, 1974), and investigated the amenability and potential contribution of BES to areas of evaluation other than job performance (Motowidlo and Borman, 1977).

Hence, although BES have failed thus far to demonstrate unequivocal psychometric superiority in reducing rating errors, the inherent, desirable characteristics of the procedure (e.g., dimensionalizing multidimensional areas, performing job analyses) have undoubtedly been a major determinant of its persistent popularity. Furthermore, Schwab, et al. (1975) noted that almost all of the comparative studies performed prior to their review involved comparisons of BES with an alternative rating scale (typically graphic) that also enjoyed the advantages of the procedures for developing BES. This is also the case for much of the comparative research to date. Thus, little is known about the value of BES as an overall procedure relative to rating scale formats that are developed from other procedures.

Mixed Standard Scales (MSS)

Mixed standard scales (MSS) were first developed in Finland by Blanz (Note 3) and later introduced into the United States by Blanz and Ghiselli (1972) as a new system for evaluating job performance. Based on analyses of leniency (mean dimension ratings), halo (interpretable factors in a factor analysis), and the relationship of the emergent factors from the factor analysis with other less merit-based variables (e.g., age, education), they suggested that this system's psychometric performance equaled and possibly surpassed that of conventional systems (e.g., graphic scales).

One need only follow the first four steps listed in Appendix I for developing BES to obtain the necessary behavioral statements. At this point, three statements are taken from each dimension. One statement exemplifies superior performance, one exemplifies moderate performance, and one depicts inferior performance, as determined by their scaled mean ratings. The statements from all dimensions are then listed together in random order, and each rater indicates whether he/she considers a ratee's performance to be better than, worse than, or similar to the behavior described in each statement.

Potential Advantages of MSS

First, Saal and Landy (1977) suggested that MSS should reduce leniency error "by forcing the rater to respond to all three descriptions of job performance for each trait being rated, as opposed to asking him to indicate which one of three descriptions best represents the ratee, and by disguising the order-of-merit of each set of three descriptions..." (p. 21).

Second, MSS lessen the abstraction in evaluations by directing attention to specific behaviors and eliminating any references to general dimensions. The focus is on evaluation of observed behaviors relative to concrete descriptions of behavior only, thereby avoiding possible contamination from inconsistent rater conceptions of performance standards engendered by ambiguous dimension labels. The word "Initiative", for example, may be interpreted differently by different individuals. Eliminating such labels might well minimize inconsistent interpretations.

Third, raters are not required to constantly pinpoint the level of performance as required on a graphic scale, but simply indicate whether the individual can be expected to perform better than, worse than, or similar (although this resembles "pinpointing") to each behavioral description.

Fourth, it allows one to observe the extent to which 1) each rater provides consistent ratings (in a Guttman sense), 2) each ratee receives consistent ratings, and 3) each dimension exhibits consistent ratings. (Consistent in this case does not imply interrater agreement on evaluated performance.) A rater, for example, may mark ratees more favorably on a statement exemplifying exceptionally good performance and less favorably on a statement exemplifying moderate performance within a certain dimension/ category. If other raters are rating the statements in a consistent manner, an investigation could attempt to determine the cause of this particular rater's apparently inappropriate ratings, and measures could then be taken to rectify the problem, if possible and desirable. If a sizable number of raters, however, are making the same inconsistent ratings on the performance category, this would necessitate a re-evaluation of the suitability of the statements representing the three performance levels.

MSS are also helpful in identifying a ratee whose performance is difficult to evaluate by revealing cases where there are generally inconsistent ratings of a ratee for certain performance dimensions. Once detected, an investigation could attempt to identify the difficulty in evaluating the individual with respect to those performance areas.

Potential Drawbacks of MSS

First, it is assumed, as with BES, that a given performance level can be represented by a single behavior and that raters are able to infer this behavior from observed behaviors. Such assumptions may indeed be dubious. The potential drawbacks of BES regarding selection criteria for behavioral statements and ambiguous ratee behavior are also germane here as well.

Second, relying upon three behavioral statements to define a dimension encompassing a multitude of behaviors may be unreasonable. Any given dimension entails behaviors of varying complexity, some of which are surely as difficult to describe as they are to evaluate. Behavioral descriptions with substantial interrater agreement on the level of performance exemplified may typically be those that are relatively easy and quick to describe, as more complex but equally important job behaviors are likely to require more time and effort and possibly some degree of rater interaction (2 or 3 raters working together) in order to be effectively delineated. Performance evaluated relative to simplistic behavioral statements does not necessarily reflect performance on more intricate job requirements. Hence, evaluations relative to simplistic statements may be accurate, but such statements might not sufficiently measure the dimension they purportedly represent. If raters focus on observed behaviors with levels of complexity similar to that described in the behavioral statements, several dimensions in MSS may be suffering some degree of criterion deficiency. A particular performance dimension might encompass other complex behaviors in comparison to which the ratee may not fare as well.

Third, previously mentioned as a potentional advantage was that a rater using MSS is not sure of the kind of evaluation given each ratee, as performance categories are not apparent and the behavioral statements are in random order, thereby disguising the order-of-merit continuum for each performance dimension. On the other hand, if raters are strongly opposed to such scales, lack of rater acceptance would assuredly jeopardize their effectiveness as a performance measure and thus attenuate their contribution to improved organizational effectiveness, as a rater's attitude and motivation undoubtedly affect the quality of the performance evaluations.

Research on MSS

Saal and Landy (1977) reported that MSS for rating police officers showed less halo error (dimension intercorrelations) and leniency error (mean dimension ratings) than BES, but also demonstrated less interrater agreement (intraclass correlation coefficients). Saal (1979) compared the relative effects of MSS and GRS for police officers and again found that MSS provided less halo (dimension intercorrelations), slightly less leniency (mean dimension ratings), and less interrater agreement (intraclass correlation coefficients).

Finley et al. (1977) developed MSS and BES for evaluating the performance of department store managers. The results suggested that BES demonstrated greater interrater agreement (multitrait-multirater analyses). Evidence for differences with respect to leniency (mean dimension ratings) and halo (multitrait-multirater analyses, factor analyses) was inconclusive.

Another recent study by Saal (Note 4) involved the development of BES and MSS for evaluating the performance of police patrol officers. Two forms of each format were developed. One set of BES and MSS was developed by supervisory officers (supervisory scales) who were currently evaluating patrol officer performance, while a second set was developed by subordinate officers (subordinate scales) who were being evaluated on patrol officer performance.

Several indices were used to measure each of the evaluated rating errors. Three different indices were used to evaluate leniency/severity: (1) departure of mean dimension ratings from the midpoint of the rating scale; (2) for each dimension, examination of the magnitude and direction of skewness in a frequency distribution of the ratings; and (3) a significant Rater main effect in a Rater x Dimension analysis of variance. Two different indices were used to evaluate halo: (1) dimension intercorrelations and (2) for each ratee, examination of the variance (standard deviation) across performance dimensions in each rater's ratings. Central tendency/ range restriction was examined using two different indices: (1) for each dimension, examination of the kurtosis in a frequency distribution of the ratings and (2) for each dimension, examination of the variance (standard deviation) across ratees in each rater's ratings. The results of these analyses suggested that "supervisory" MSS were less contaminated by leniency, halo, and central tendency errors than "supervisory" BES. No reliable differences were found between "subordinate" BES and "subordinate" MSS.

It appears that, relative to BES and GRS, MSS have generally been less contaminated by halo and leniency errors and have generally exhibited less interrater agreement.

Training Raters To Minimize Rating Errors

The implication of the comparative studies examining the relative psychometric properties of BES and MSS is that format modifications and improved scale development procedures do not ensure the emergence of improved performance ratings with respect to such desirable psychometric properties as independence of dimension ratings, differentiation of ratee perfromance on each dimension, interrater agreement on ratee performance, and effective use of the full range of the scale. Consequently, several investigators have alternatively proposed that attention should be directed toward training raters to minimize rating errors (Borman and Dunnette, 1975; Burnaska and Hollmann, 1974), as individuals who are delegated the uncomfortable responsibility of evaluating the performance of others may simply not command sufficient knowledge or skill to avoid undesirable response tendencies, regardless of format and scale development procedures.

Potential Advantages of Rater Training

First, it is possible that an acute awareness by raters of errors that attenuate the validity of performance ratings may engender relatively more careful and objective judgments exhibiting improved statistical properties. Although rater bias may be immune to total suppression, in that raters might not totally overcome idiosyncratic response tendencies or completely divorce personal feelings from perceptions of ratee behavior, awareness of contaminating influences may significantly reduce their impact on the performance ratings.

Another potential advantage is that if raters are trained prior to observing ratee performance, they may be more inclined to begin to constantly perceive and evaluate performance in a multidimensional manner. That is, such training may actually enhance a rater's cognitive complexity, as opposed to inducing only a temporary change. This would indeed be desirable, as Schneier (1977) found that cognitively complex raters exhibited less halo than cognitively simple raters, regardless of the complexity of the scale format.

Finally, at an intuitive level it would seem that helping raters to be as objective and perceptive as possible and making them acutely aware of the potential pitfalls (e.g., halo, leniency/severity, restriction of range, etc.) in the rating situation would be beneficial for improving their effectiveness as judges. Training is usually not only valuable, but also a requirement for many jobs, and this would appear to hold for performance appraisal as well, especially in light of ever increasing rater accountability due to greater federal scrutiny of merit-based selection and promotion ratings. Apart from the psychometric advantages, training may benefit the raters by increasing their confidence in their ability to evaluate others simply by knowing what to look for as well as what to look out for, thereby possibly lessening the anxiety that generally accompanies performance evaluations. Raters may also adopt a more behavioral approach to discussing the evaluations with their subordinates, which would benefit both the raters and the ratees.

Potential Drawbacks of Rater Training

The major determinant of rating quality is rating accuracy (validity), regardless of the relative psychometric worth of the ratings. Unfortunately, a "true" measure of rating error (rating inaccuracy) is not feasible in the real world of performance ratings, as a "true" performance rating does not exist. There is, however, a concern that increased differentiation of dimensions and ratees which is induced by rater training may adversely affect the validity of ratings as well. Although untestable (except under highly artificial conditions involving ratings of vignettes), increased variance in raters' ratings may be in response to the trainer's underlying request, but not necessarily an indication that the evaluations are any more accurate. It is hoped that the overall effect of such amplification is an increase in accuracy. On the other hand, it is conceivable that training may compel raters to make several amplified distinctions among the ratees and dimensions that they are really not sure of and that may well be unjustified. Thus, the overall impact of training on rating accuracy is not altogether clear.

Research on Training

Latham, Wexley and Pursell (1975) conducted a study wherein managers of a large coproration evaluated hypyothetical job applicants observed on videotape. Sixty managers were randomly subdivided into three groups of equal size. Two groups participated in either a workshop or discussion session involving training directed toward reducing rating errors, whereas the control group received no training. Six months later the managers evaluated the hypothetical job applicants. While the control group was guilty of halo error, both experimental groups were successful in reducing this type of bias. Halo was measured by providing one subset of a rater group with folders containing extremely unfavorable information on an individual and another subset of that rater group with very favorable information on the individual. No information, however, pertained to educational background. Managers rated the individual on education and other job-related factors. If the ratings on education by raters given unfavorable information were significantly lower than the ratings on education by those given favorable information, halo error was thought to be present.

Borman (1975) had "low" and "middle" managers rate vignettes (developed from previously scaled behavioral statements) describing the performance of first-line supervisors. Managers used BES to rate the vignettes prior to, and immediately following, a five minute training session, wherein halo error was described and raters were urged to avoid it. While the results indicated that training reduced halo error (variance of each rater's ratings of a ratee across dimensions relative to the ratee's "true" variance), validity (for each dimension, mean ratings for the vignettes were correlated with the corresponding "true" cirterion scores of the vignettes) of the ratings remained close to pretraining levels, and an attenuation of interrater agreement (intraclass correlation coefficients) occurred.

Bernardin and Walter (1977) developed BES for rating instructors and compared group differences as a function of amount of training, amount of exposure to BES, and the proximity (i.e., beginning vs. end of semester) of training and exposure to the performance ratings, which were performed at the close of the semester. They reported that any group with more training, more exposure to BES, and/or earlier training/exposure to BES consistently demonstrated less halo error (standard deviation of each rater's ratings across dimension). As for leniency (ratee means compared in Group x Dimension x Ratee ANOVA), the most highly trained and "exposed" group showed less leniency error than all other groups, and the group differing from the most highly trained and "exposed" group only with respect to exposure to BES showed less leniency than the untrained group. No differences were found with respect to interrater agreement (standard deviation of ratings on each dimension for each ratee) between the untrained group and any trained group. Although the finding concerning interrater agreement is inconsistent with Borman's (1975) results, the substantive differences between their designs (ratings of vignettes vs. ratings of "real" people) should be noted, including the disparate operational definitions.

Bernardin (1978) investigated the effects of different amounts of training over time. One group of students received one hour of training:

> ". . . as in Bernardin and Walter (1977). This training involved definitions, graphic illustrations, and examples of the errors of leniency, halo effects, and central tendency. Students were also given data to evaluate in terms of the errors, and the evaluations were discussed. Reference was made to the several performance dimensions subsequently measured on the rating instruments (BES and SRS), but students were not shown the actual scales. Members of Group 2 were given a 5-minute training session (B) on error similar to that described by Borman (1975). Definitions of the three errors (leniency, halo effect, and central tendency) were presented, and one graphic illustration was presented of each. Reference was also made to the dimensions on the rating instruments." (p. 302).

An interesting finding that emerged from this study was that subsequent reductions in psychometric error resulting from training returned to pretraining levels after a twelve week interval. Apparently, some form of iterative training may be necessary.

A study by Vance, Kuhnert, and Farr (1978) examined the effects of rater training and scale format on the psychometric quality of interview ratings. College students used either behaviorally anchored scales or graphic scales (behavioral anchors were eliminated) to evaluate vignettes of hypothetical candidates applying for the job of resident assistant in a college dormitory. As in Borman (1975), the values of the previously scaled behavioral statements provided external performance criteria for the examination of several rating errors. The findings showed that the ratings by students who used behaviorally anchored scales were closer to the "true" performance levels, less contaminated by halo and leniency errors, and higher in interrater agreement (intraclass correlations). Each of the analyses for rating errors involved scores derived from the absolute difference between the "true" scores for each rating error and the raters' scores. The results also showed, however, that a brief training session designed to familiarize raters with common rating errors and urge raters to avoid such errors had no effect on the psychometric quality of the ratings.

Borman (1979) examined the effects of rater training and scale format on the psychometric quality of performance ratings. College students in the trained and untrained groups used one of five different formats (BES, behavior summary, summated, trait, and GRS) to rate the videotaped performance of ratees who acted out scripts exemplifying various performance levels of managers and recruiters. Rater training basically consisted of a 3-hour training session on rating errors designed by Latham et al. (1975). The results indicated that (1) summated rating scales were noticeably less contaminated by halo error (dimension intercorrelations for each rater, Ratee x Dimension interaction measure); (2) there was only weak evidence to suggest that training effectively reduced halo for any scale format; (3) with the possible exception of relatively low convergent validity for the trait rating scale, no training or format differences were readily apparent regarding convergent and discriminant validity (modified version of the ANOVA method by Kavanagh, MacKinney, and Wolins, 1971); and (4) the obtained results did not identify one format as consistently better than any other format with respect to rating accuracy (for each rater, ratings of the eight ratees were correlated with the mean "true" scores on a given dimension, and the 2 scores for the coefficients were entered into Format x Training x Job and Format x Dimension x Training ANOVAs).

Purpose of Study

Nearly all research concerned with the psychometric quality of ratings has focused on either scale format or rater training. While research has failed to unequivocally demonstrate inherent psychometric strengths and weaknesses in different rating scale formats, studies examining rater training have demonstrated that common rating errors (i.e., halo and leniency) can be reduced, at least temporarily, when raters are trained to avoid them. The effectiveness of training raters to improve their ratings, however, may in fact be partially determined by the scale format. Such differential training effects may further demonstrate that the relative psychometric properties of rating scale formats are not consistent between ratings by untrained raters and ratings by trained raters. A brief training session could conceivably nullify, amplify, or in some way modify findings regarding common rating errors that originally emerged from ratings by untrained raters. Such interactions of training and format would have important implications for interpreting available research and for subsequent investigations designed to improve the performance ratings of raters.

Thus, two related objectives of this study were to (1) examine the possible impact of scale format on the effectiveness of a brief training session for raters and (2) determine whether or not the relative psychometric properties of different formats when using untrained raters are consistent with the relative psychometric properties of different formats emerging from ratings by trained raters.

In light of the hypothesized advantages mentioned in the review and the heightened popularity, behavioral expectation scales (BES) and mixed standard scales (MSS) were therefore selected as the formats to be examined. Conventional graphic scales (GRS) were also included as a kind of standard format against which to compare ratings on BES and MSS. The training session and the procedures for developing each scale are described later.

A secondary goal of this study was the development of two forms of BES and two forms of MSS. Both forms of each scale were included in the study in order to determine not only whether different formats, but also whether identical formats with qualitatively different behavioral statements (corresponding statements had similar scale values) exhibit disparate psychometric properties.

Zedeck, Jacobs and Kafry (1976) developed alternate forms of BES using statements generated in the Harari and Zedeck (1973) study and reported very favorable results. It should be noted, however, that they were not making appropriate tests of the criteria for parallel tests discussed by Ghiselli (1964). In performance appraisal, raters are analogous to test items. A ratee's "test score" may be calculated by summing or averaging the raters' ratings (a separate test score for each dimension). Thus, Zedeck, et al. (1976) used a 44-item test form (44 student raters used form A) and a 51-item test form (51 student raters used form B), but only one individual (the instructor) was evaluated on the two forms of the test (BES). The criteria of parallel tests pertain to the scores of a sample of individuals evaluated by both forms, not to the distribution of ratings (test items) on each form.

If significant psychometric differences are found between rating scales that differ with respect to the behavioral descriptions only, this would, for MSS, raise serious questions concerning the comparability of the corresponding behavioral descriptions of the alternate forms that, according to the scale development procedure, should exemplify essentially the same level of performance for the same dimension. For BES, psychometric differences would suggest that the anchors for each form are eliciting different responses from the raters, which should not occur if each behavior on each form was accurately exemplifying the specified area of performance and the specified level of performance.

Method

This study can be subdivided into two phases. The first phase included the development of three rating scale formats to evaluate instructor performance, while the second phase involved the evaluation of instructor performance by students who used different scale formats and participated in different training conditions. The data from phase II was used to evaluate training and format differences. Each phase is described below.

Phase I (Scale Development)

Development of BES

BES for evaluating college instructors were developed according to the five basic steps delineated in Appendix I. (The procedures used in this study are specified below.) A total of 147 college students from four general psychology classes participated in the scale development phase, each of whom was satisfying experiment-participation requirements. Each step outlined below was preceded by a brief review of the developmental procedures and a rationale for using the Smith and Kendall (1963) approach. Students were allowed to participate in only one meeting or step in the developmental phase.

Step 1. The first step was modified to include two groups (N=20 in each group). Participants in the first group were asked to identify by means of a single word or short phrase all dimensions of teaching that they believed were important and could be satisfactorily evaluated by students. Each dimension was listed on a blackboard. The students were subsequently instructed to define in behavioral terms each of the identified dimensions. Each student was provided with a pencil and scratch paper for this task. Apparently redundant dimensions were either eliminated or combined into a single dimension at this time. The second group reviewed the dimensions for clarity and made any revisions deemed necessary. Nine dimensions emerged from this procedure.

Step 2. Students (N=39) were given a booklet containing a separate dimension on each page and were asked to describe three instructor behaviors for each dimension. One behavior was to exemplify exceptionally good performance, one moderate performance, and one exceptionally poor performance. Although 1,053 behavioral statements were initially generated, elimination of repetitive or ambiguous statements by the author reduced the number to 331. Minimal editing of grammatical errors was also performed by the author at this point.

Step 3. Students (N=36) were provided with a list of the nine dimensions and a sheet of paper containing 331 numerically labeled spaces. The author read aloud each numeral and corresponding behavioral statement that had been randomly assigned to the numeral. Each student then assigned to the similarly numbered space on the sheet of paper the label of the dimension to which the statement "belonged." The criteria for subsequent selection of behavioral statements were that (1) a statement must be assigned to the same dimension by at least 75% of the students and (2) assignment to any other single dimension by the students was not to exceed 20%. The number of statements meeting or surpassing these criteria was 192.

Step 4. Students (N=32) were given a booklet containing the 192 statements that had survived the retranslation criteria and instructed to rate each statement according to the level of teaching performance it exemplified in relation to the dimension to which it was assigned. The statements were presented in the form of behavioral expectations in order that ratings would be assigned to the statements in the form that they would appear on the scales, thereby avoiding any possible contamination from post hoc modifications of the statements into the form of expectations (see Appendix II).

Step 5. A total of 96 statements were selected from the 192 scaled statements to develop the two forms of BES. The criteria for inclusion in the scales were that (1) the statement must have a scaling standard deviation below 1.5 (Two statements, however, with standard deviations slightly above 1.5 were included as anchors on one dimension in order to fill an otherwise large "gap."); (2) each dimension must have only one statement per interval on a 7-point scale (six statements maximum per dimension, as the intervals were 1-2, 2-3, 3-4, etc.); and

(3) the anchors on a scale should describe qualitatively different types of behavior rather than describe the same type of behavior at various levels of performance. The aforementioned procedures produced two forms of BES designed to measure nine job performance dimensions: assignments; attitude towards subject; grades; instructor knowledge; manner of presentation; objectiveness; organization; student-teacher relations; and tests. The two forms of BES are presented as Appendix III.

The certainty ratings accompanying each dimension of the BES were included for purposes unrelated to the objectives of this study. This was true for MSS and GRS as well. It should therefore be noted that certainty ratings did not enter into any of the analyses.

Development of MSS

The highest anchor, lowest anchor, and the anchor with a mean value closest to 4.0 were selected from each dimension of Form A of the BES to construct Form A of the MSS. Thus, MSS consisted of 27 statements (9 dimensions x 3 statements per dimension) listed in random order. Form B of the MSS was constructed in a similar fashion, including use of the same random ordering so as to avoid potential differences engendered by different random orderings of the statements. The two forms of the MSS, along with an explanation of the rating procedure, are presented as Appendix IV.

Two systems for numerically coding the response combinations were used in the subsequent analyses of the ratings on MSS. Saal (1979) introduced an alternative coding system that warrants strong consideration on the basis of greatly improved consistency in the coding of response combinations. The numerical values of the response combinations for both the original (Blanz & Ghiselli, 1972) and revised coding systems, along with a description of the revised system, are presented as Appendix V.

Development of GRS

GRS were developed by simply replacing the behavioral anchors on the BES with traditional adjective anchors. That is, the top part of each scale was labeled "Exceptionally Good," whereas the bottom part was labeled "Exceptionally Poor." Thus, GRS were afforded all the advantages inherent in the developmental procedures of BES, less the inclusion of behavioral anchors (see Appendix VI).

Phase II (Rater Training and Performance Ratings)

Performance Evaluations

During the fourteenth week of the spring semester students in each of two general psychology classes at Kansas State University were invited to evaluate their instructor, and volunteers (N=206) were awarded experiment-participation credit. Participants in each class were randomly subdivided into two groups. One group remained in the room for a 20-minute training session prior to rating the instructor, while the second group was escorted to a different room to evaluate the instructor without prior training.

One of the five rating scales (form A or B of BES, form A or B of MSS, or GRS) was randomly distributed to each student. Subsequent to the distribution of the rating scales, students were instructed to read the directions accompanying their respective forms before proceeding to evaluate the instructor. The number of different forms distributed to each group is reported in Table 1. Performance ratings were obtained from each class at their respective meeting times.

Table 1

Number of Raters in Each Experimental Group

	BES		MSS		GRS
	Form A	Form B	Form A	Form B	
		Trained Raters			
Instructor I	14	13	15	15	15
Instructor II	11	11	10	10	11
		Untrained Raters			
Instructor I	12	9	11	11	13
Instructor II	5	5	5	5	5

The considerable difference in the number of raters for trained and untrained groups resulted from an unexpectedly high attrition rate. Many students in the untrained groups chose to leave the experiment while walking to the other room. Such a subject loss possibly produced rater characteristic differences between training groups concerning attitude and motivation, which may have confounded the results of the experiment. Students in the untrained groups, for example, may have had better attitudes and been more motivated, since uninterested students in the untrained groups could leave if they so desired. It is likely that students who were more willing to participate in the experiment did a better job in attempting to accurately rate the instructor. It should be noted, however, that all of the students in the experiment had initially volunteered to participate.

Training

A pamphlet containing a brief description of halo, leniency/severity, and the tendency to assign only moderate ratings (see Appendix VII) was distributed to each student in the training groups. The author reviewed the pamphlet with the students and encouraged them to share any questions or comments with the group. Discussions ensued between and among the students and author regarding the effects of such errors on the accuracy and subsequent value of the ratings in providing feedback for improving performance and assisting administrative decision making. At the conclusion of each training session, students were urged to avoid these response tendencies when evaluating the instructor.

Analyses

Leniency/Severity. For each instructor, differences in mean dimension ratings of students who participated in different training conditions or used different rating scales were examined through a series of Training x Rating Scale multivariate analyses of variance (MANOVA). The nine performance dimensions were the dependent variables. Separate examinations of each dimension were performed following the emergence of a statistically significant mean vector score difference in order to determine whether or not the mean differences for the dimensions were in a common direction. No attempt was made to interpret significant multivariate differences when the directions of the univariate mean differences between rating scales or training conditions were not in one general direction. The intent was to identify significant multivariate differences wherein the univariate differences were in one readily discernable direction. This would suggest that raters in a particular group gave generally higher/lower ratings to the instructor.

The significance of a multivariate mean difference is determined by the covariance relationships among the dimensions, as well as by the means and variances associated with each separate dimension. Furthermore, a significant multivariate mean difference can be entirely attributable to a significant difference on one dimension alone, with no other dimensions exhibiting any mean differences whatsoever. The concern here, however, is not with covariance relationships or with significant multivariate differences that result from a small number of univariate differences, but rather with differences which indicate that dimension ratings associated with a particular training condition or rating scale are indeed generally higher/lower. Hence, to advocate the presence of "level" differences, the dimensions should clearly show differences in a common direction.

As an additional aid to interpreting the significance of differences in mean dimension ratings between specific rating groups, the randomization test for matched pairs (Siegel, 1956) was also performed following the emergence of a statistically significant MANOVA. The difference between the means of the two groups on each of the nine dimensions constituted the nine difference scores to be entered into the analysis. It should be noted that the difference scores were not independent estimates of a general "level" difference, as the same raters performed the ratings on each dimension. Assuming, however, that random sampling has sufficiently suppressed group differences in mean ratings resulting from effects due to other than the experimental variable alone, the analysis can be expected to help disclose whether or not the mean ratings of a particular rating scale or training group tended to be consistently higher/lower across the performance dimensions.

Halo. The index used to examine halo was the variance in each rater's ratings of the nine performance dimensions. A small variance suggests that the student is not differentiating the different performance areas of the instructor, thus indicating the presence of halo. A large variance, on the other hand, suggests that the student is perceiving differences in the instructor's respective levels of performance on each dimension, thus indicating the relative absence of halo.

A series of two-factor and single-factor analyses of variance (ANOVA) were performed as follows: First, the variance scores of all raters in the groups being examined in the particular analysis were rank ordered. That is, the variance scores of all raters included in the analysis were ranked in a single series, regardless of group membership. Rank scores were then substituted for the corresponding variance scores. An analysis of variance was performed on the groups' rank scores to determine the significance of rank score differences between the groups. Kemp and Dayton (Note 5) demonstrated that the Kruskal-Wallis nonparametric single-factor analysis of variance by ranks can be replaced by a parametric t or F test of the rank scores, as the analyses provide essentially indistinguishable results. Although the two-factor ANOVAs of rank scores may be somewhat questionable, they were simply used as initial tests of the overall effects of Training and Scale Format for each instructor.

One of the major tasks of any researcher is to select appropriate statistical models for analyzing the obtained data. The statistical tests of significance in ANOVA were derived from a mathematical model that is predicated on fundamental theoretical notions concerning the population(s) being sampled. Such assumptions, however, may be untenable for populations consisting of variance or standard deviation scores. If the assumptions cannot be upheld, then the interpretability of the significance tests is obscured, as the mathematical model from which the tests were derived is incompatible with the population being sampled. A $p < .05$ finding, for example, may differ considerably from what it should "really" be. Since use of ANOVA for analyzing variance scores has not been justified, a more conservative nonparametric test was used instead. Again, the reason for using ANOVA to analyze the rank scores was because such results have been found to be consistent with findings that would have otherwise emerged from the Kruskal-Wallis nonparametric test.

Interrater Agreement. The index used to examine interrater agreement was the variance in the ratings by students who were in the same training condition, used the same rating scale, and evaluated the same instructor. A variance was calculated for each dimension. A small variance suggests that the students are giving similar ratings, thus indicating the presence of interrater agreement, whereas a large variance suggests that the students are not giving similar ratings, thus indicating the relative absence of interrater agreement.

First to be performed were a series of multivariate tests of homogeneity of the within variance/covariance matrices (1) between training groups who used the same rating scale and evaluated the same instructor and (2) between two rating scales, while holding both Training and Instructor constant. A statistically significant finding would suggest that the two groups possessed different variances and/or covariances. If a statistically significant difference was found, each of the nine variance components, constituting the trace (diagonal) of the matrix, was compared to the corresponding variance component of the other matrix. (At this time, there is no straightforward statistical test which discloses the significance of trace differences alone.) For each dimension, an _F_ ratio was derived by dividing the larger variance by the smaller variance. According to the respective degrees of freedom for each variance, the statistical significance of the _F_ value was then identified in an _F_ table. A nonsignificant variance ratio _F_ would suggest that the two variances might well have arisen from random samples of populations with equal variances. Again, the intent was to identify significant multivariate homogeneity differences that were clearly attributable to generally larger dimension variances for a particular group.

To perform a two-tail test of significance, the _p_ values had to be interpreted differently, as the specified _p_ values pertain to one-tail tests. Since the larger variance is arbitrarily placed in the numerator, the probability of obtaining deviations _above_ the mean _F_ is doubled. Therefore, in order to perform a two-tail test of the variance ratio _F_, the probabilities in the _F_ table were doubled (i.e., .05 became .10, .025 became .05, etc.).

No analyses were performed with the individual covariances, as covariance relationships are not relevant to the assessment of interrater agreement.

As an additional aid to interpreting the significance of variance differences between specific rating groups with respect to interrater agreement, the Wilcoxon matched-pairs signed-ranks test (Siegel, 1956) was also performed either following the emergence of a significant multivariate test of homogeneity or whenever the homogeneity test could not be performed due to matrix singularity. (The randomization test was not used, since an interval scale for variance scores was not assumed.) The difference between the variances of the two groups on each of the nine dimensions constituted the nine difference scores to be entered into the analysis. As with the randomization test, the difference scores were not independent estimates of a general difference in interrater agreement. Assuming again that random sampling has satisfactorily equated the groups, the analysis can be expected to help show whether or not the variances of a particular rating scale or training group tended to be consistently larger/smaller across the performance dimensions.

Results

The following results involved a sizable number of analyses. Therefore, in light of a greatly increased experimentwise error rate, the author chose to adopt a higher level of confidence ($p \leq .01$) so that group differences would be at a level which would ensure a certain reliability of the conclusions.

Leniency/Severity

Table 2 shows the mean dimension ratings and standard deviations for each rating scale, which are subdivided according to instructor and training condition.

For each instructor, a set of multivariate analyses of variance (MANOVA) involving all five rating scales and both training conditions were first performed to determine whether or not there were any significant overall differences in the mean ratings. Two analyses were performed for each instructor. One analysis included the original coding system for MSS, while the other included the revised coding system. The results reported in Table 3 indicate the presence of a significant Rating Scale ($p < .001$) effect for Instructor I, whereas there were no significant differences found for Instructor II.

An examination of the univariate analyses for Instructor I revealed five dimensions with significant Rating Scale effects that appeared with both the original and revised MSS coding systems: attitude towards subject; instructor knowledge; manner of presentation; objectiveness; and

Table 2

Mean Dimension Ratings and Standard Deviations for BES, MSS, and GRS

Dimension	BES				MSS (original)				MSS (revised)				GRS	
	Form A		Form B		Form A		Form B		Form A		Form B			
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
	Untrained raters for instructor I													
Assignments	5.92	.88	5.83	.79	5.91	.70	5.18	.75	5.91	.70	5.18	.87	5.77	1.13
Attitude Towards Subject	6.42	.79	6.39	1.02	5.18	1.17	5.18	1.78	5.36	.92	5.55	1.57	6.65	.38
Grades	5.88	1.17	5.50	1.06	4.82	1.17	5.27	.91	4.91	.94	4.73	1.01	6.00	.76
Instructor Knowledge	6.00	.91	5.89	.82	4.55	.69	5.36	1.36	4.82	.75	5.55	1.21	5.96	1.35
Manner of Presentation	6.38	.86	6.28	.76	4.36	1.43	5.09	2.02	5.00	1.00	5.09	1.30	6.23	.56
Objectiveness	5.63	1.13	6.22	.62	5.36	.67	5.73	1.10	5.36	.67	5.46	1.04	5.96	1.05
Organization	6.29	.62	6.17	.71	5.64	.51	5.73	1.10	5.64	.51	5.46	1.37	6.04	.122
Student-Teacher Relations	5.46	1.31	5.61	1.41	4.64	.67	5.18	.60	4.64	.67	5.00	.89	5.58	1.46
Tests	4.79	1.60	3.78	1.50	4.91	1.22	5.09	1.30	5.09	1.04	5.27	.79	5.19	1.97

Table 2 (Cont)

Dimension	BES				MSS (original)				MSS (revised)				GRS	
	Form A		Form B		Form A		Form B		Form A		Form B			
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
	Trained raters for instructor I													
Assignments	5.68	.93	5.50	1.38	5.33	.90	4.53	.83	5.20	.94	4.73	.96	5.90	1.30
Attitude Toward Subject	6.29	.73	6.42	.89	5.13	.92	5.93	.96	5.27	.59	5.60	1.24	5.67	1.08
Grades	5.46	1.53	4.96	.97	4.80	1.21	4.40	1.06	4.53	1.30	4.27	1.03	5.17	1.71
Instructor Knowledge	6.18	.64	5.73	.81	4.00	1.36	5.60	.99	4.40	1.06	5.60	.99	5.37	.72
Manner of Presentation	5.46	1.77	5.69	1.33	5.00	1.13	5.20	1.21	4.80	.94	5.33	1.18	5.43	1.51
Objectiveness	6.00	.62	6.31	.66	5.00	.76	4.87	1.36	4.87	.92	4.93	1.39	5.40	1.27
Organization	6.32	.46	5.00	1.48	4.87	1.25	4.93	.88	4.80	1.01	5.23	.64	5.00	1.57
Student-Teacher Relations	4.32	1.17	5.15	1.60	5.07	.88	5.07	1.10	5.00	.76	4.93	1.16	5.00	1.84
Tests	4.43	1.16	4.09	1.90	4.47	1.30	4.23	1.27	4.47	1.19	3.87	10.6	3.73	1.44

Table 2 (Cont)

Dimension	BES				MSS (original)				MSS (revised)				GRS	
	Form A		Form B		Form A		Form B		Form A		Form B			
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
	Untrained raters for instructor II													
Assignments	4.90	1.03	5.70	1.44	5.80	.45	4.80	.84	5.80	.45	5.20	.84	5.80	.45
Attitude Toward Subject	5.80	1.04	5.30	1.10	5.60	.55	5.80	.45	5.60	.55	5.80	.45	5.70	1.10
Grades	5.40	1.14	4.20	.84	5.40	.55	5.20	1.30	5.40	.55	5.20	1.30	5.20	.91
Instructor Knowledge	6.10	.89	4.50	1.70	5.00	.00	5.60	.55	4.80	.45	5.60	.55	5.10	1.25
Manner of Presentation	4.10	1.34	3.90	.89	5.00	.00	4.20	1.30	5.00	.00	4.20	.84	4.00	.71
Objectiveness	6.00	1.70	4.80	1.48	5.00	1.00	5.00	1.23	5.00	1.00	4.80	1.64	4.40	.42
Organization	5.60	.89	6.30	1.04	5.80	.45	4.60	1.14	5.60	.89	4.40	1.52	5.92	.55
Student-Teacher Relations	5.80	1.44	5.90	1.39	5.40	1.14	5.20	.84	5.80	.84	5.00	1.23	5.20	.57
Tests	3.50	1.46	4.90	1.85	5.20	1.10	4.40	2.07	5.20	1.10	4.20	1.92	4.00	1.23

Table 2 (Cont)

Dimension	BES				MSS (original)				MSS (revised)				GRS	
	Form A		Form B		Form A		Form B		Form A		Form B			
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
	Trained raters for instructor II													
Assignments	5.27	1.27	5.41	1.11	5.10	.99	5.00	1.16	5.30	.68	5.00	1.16	4.77	1.68
Attitude Toward Subject	5.77	.82	6.00	.87	5.70	.95	5.70	1.25	5.90	.74	5.90	.88	6.55	.52
Grades	5.32	1.38	5.14	1.00	4.70	1.42	5.00	1.33	4.70	1.34	5.10	1.52	5.46	1.51
Instructor Knowledge	5.59	1.07	5.86	.95	4.80	1.03	5.00	1.76	4.80	1.14	5.00	1.76	5.18	.87
Manner of Presentation	4.59	1.22	5.05	1.15	3.80	.92	4.30	1.16	4.20	1.32	4.40	1.17	5.09	1.18
Objectiveness	5.05	1.25	6.09	.49	5.00	.67	4.80	1.03	5.00	.67	4.70	1.25	4.73	1.57
Organization	5.68	1.60	6.46	.47	4.80	1.14	5.10	1.20	4.90	.88	5.10	1.10	5.59	1.34
Student-Teacher Relations	5.50	1.61	5.64	1.25	5.50	.97	5.10	.88	5.60	.84	5.10	1.10	5.23	1.54
Tests	4.91	1.51	4.73	1.57	4.20	1.40	4.40	1.51	4.20	1.32	4.40	1.43	4.23	1.54

Note: BES = Behavioral Expectation Scales
MSS = Mixed Standard Scales
GRS = Graphic Rating Scales

Table 3

Multivariate Analyses of Variance (Leniency/Severity)

for Rating Scales and Training Conditions

MANOVA				
Sources of				
Variance	Instructor I[a]	Instructor II[a]	Instructor I[b]	Instructor II[b]
Rating Scale	F(36,413) = 3.13*	F(36,226) = 1.42	F(36,413) = 2.88*	F(36,226) = 1.48
Training	F(9,110) = 2.20	F(9,60) = .61	F(9,110) = 1.88	F(9,60) = .96
Ra x Tr	F(36,413) = 1.59	F(36,226) = 1.24	F(36,413) = 1.21	F(36,226) = 1.07

* $p < .001$

[a] original MSS coding system was used

[b] revised MSS coding system was used

organization. In addition, the dimension labeled "assignments" was significant when the original MSS coding system was used, and the dimension labeled "grades" was significent with the revised MSS coding system.

Based on the above findings, sets of Rating Scale (2) x Training (2) MANOVAs were performed on the ratings of Instructor I, wherein all possible pairs of rating scales were compared. The first set of analyses examined form and training effects, while the second set examined format and training effects.

Form effects--The analyses revealed only one significant multivariate form effect, which was between MSS (A) and MSS (B) when the original coding system was used. The univariate analyses, however, showed only two dimensions with significant mean differences (assignments and instructor knowledge). Moreover, the differences were in opposite directions, thus suggesting that the ratings on a particular form were not generally higher/lower. The results of the randomization test further indicated no significant difference in the "levels" of the mean dimension ratings between the two forms. The results concerning training revealed no significant multivariate training effects.

Format effects--In order to maintain a certain independence (not using the same data in several related analyses) among the analyses focusing on format effects in mean dimension ratings and avoid confounding format effects with form effects, each form of BES was only compared to the corresponding MSS form. Both forms of BES and MSS were compared to GRS, however.

The results of the format analyses are reported in Table 4. Significant multivariate differences were found between BES and MSS for both Form A and Form B. An examination of the univariate components (dimensions) constituting the multivariate vectors of BES (A) and MSS (A) showed that BES (A) had significantly higher mean ratings on five dimensions. Further examination of this apparent "level" difference with a randomization test revealed significantly higher mean ratings on BES (A) for untrained raters, but not for trained raters. For Form B, however, BES had significantly higher mean ratings than MSS or revised MSS on only two or one dimension(s), respectively, thus indicating the possible absence of general format differences. The nonsignificant randomization tests further suggest that there were no general differences in leniency/severity between BES (B) and MSS (B) due to sufficiently large reversals in the direction of mean differences among the dimensions. Finally, the analyses involving GRS and MSS (A) revealed the presence of significantly higher ratings on GRS which,based on the randomization test, again seemed to be moreso for untrained raters.

Training effects--The results appearing in Table 4 indicate that there were no significant multivariate mean

Table 4

Multivariate Analyses (Leniency/Severity) of Format

Combinations for Instructor I

Rating Scale Formats						Dimensions[c] of Scale 1	Randomization Test	
Scale 1		Scale 2	T[a]	F[b]	TxF	With Higher Means	Untr[d]	Tr[e]
BES(A)	vs.	MSS(A)	ns.[f]	.001[g]	ns.	$\overset{**}{B}$, $\overset{**}{D}$, $\overset{*}{E}$, $\overset{*}{F}$, $\overset{**}{G}$	$p < .01$	ns.
BES(A)	vs.	revised MSS(A)	ns.	.001	ns.	$\overset{**}{B}$, $\overset{**}{B}$, $\overset{*}{E}$, $\overset{*}{F}$, $\overset{**}{G}$	$p < .01$	ns.
BES(B)	vs.	MSS(B)	ns.	.01	ns.	$\overset{*}{A}$, $\overset{*}{F}$	ns.	ns.
BES(B)	vs.	revised MSS(B)	ns.	.01	ns.	$\overset{**}{F}$	ns.	ns.
BES(A)	vs.	GRS	ns.	ns.	ns.			
BES(B)	vs.	GRS	ns.	ns.	ns.			
GRS	vs.	MSS(A)	ns.	.001	ns.	$\overset{**}{B}$, $\overset{**}{D}$, $\overset{*}{E}$,	$p < .01$	ns.
GRS	vs.	revised MSS(A)	ns.	.001	ns.	$\overset{**}{B}$, $\overset{**}{D}$, $\overset{*}{E}$	$p < .01$	ns.
GRS	vs.	MSS(B)	ns.	ns.	ns.			
GRS	vs.	revised MSS(B)	ns.	ns.	ns.			

[a] Training effect

[b] Format effect

[c] A=Assignments, B=Attitude Towards Subjects, C=Grades, D=Instructor Knowledge, E=Manner of Presentation, F=Objectiveness, G=Organization, H=Student-Teacher Relations, I=Tests

[d] Test performed on ratings of untrained raters

[e] Test performed on ratings of trained raters

[f] Not significant at the .01 level

[g] Level of statistical significance associated with source of variance ($p \leq$)

* $p < .01$

** $p < .001$

differences between training groups, regardless of format.

The above results therefore suggest that Instructor I received significantly higher ratings on BES (A) and GRS than on MSS (A), and that the differences appeared to be greater for untrained raters. Since this finding was neither replicated over instructors nor over forms, however, evidence of general format effects regarding leniency is inconclusive.

Halo

Single-factor analyses of variance (ANOVA) of rank scores were used to examine the effects of Training and Rating Scale on halo. The first set of analyses examined form effects, while the second set examined format effects. Both Instructor and Training were held constant for each analysis. A third set of analyses was also performed to determine the effectiveness of training with different rating scales. Both Instructor and Rating Scale were held constant for each of these analyses. Table 5 shows the mean halo variance in each rater group.

For each instructor, 3 x 2 ANOVAs (3 formats, 2 training conditions) of rank scores were initially performed to examine overall format and training effects. The results appear in Table 6. Although no differences were found with Instructor I, a significant Training x Format interaction emerged from the analyses involving Instructor II. The nature of the interaction was disclosed in the following Format and Training comparisons.

Table 5

Mean Variance Scores (Halo) of Training and Format Combinations for Each Instructor

Rating Scales	Instructor I		Instructor II	
	Untrained	Trained	Untrained	Trained
BES(A)	.89	1.39	1.80	1.07
BES(B)	1.13	1.44	1.50	1.15
GRS	1.08	1.32	.99	1.49
MSS(A)	.82	1.04	.37	1.17
revised MSS(A)	.58	.71	.43	1.02
MSS(B)	.98	1.06	.67	1.26
revised MSS(B)	.75	.93	.65	1.10

Table 6

Analyses of Variance of Rank Scores (Halo) for Training Conditions and Rating Scale Formats

Instructor I

Sources of Variance	original MSS		revised MSS	
	F	*p*	*F*	*p*
Training	1.84	.18	2.73	.10
Format	.60	.55	2.97	.06
T x F	.07	.93	.03	.97

Instructor II

Sources of Variance	original MSS		revised MSS	
	F	*p*	*F*	*p*
Training	.97	.33	.31	.58
Format	3.92	.03	6.02	<.01
T x F	6.53	<.01	5.01	<.01

Form effects--For both instructors, there were no halo differences between forms for either BES or MSS, regardless of whether trained or untrained raters performed the ratings. The effects of training on alternate forms were also examined at this time and the results further indicated a similar effect of training on halo for the alternate forms of a particular format. Thus, based on the findings that emerged from these analyses, variance scores of alternate forms were combined to form a single format group for the remaining analyses involving format and training effects.

Format effects--For Instructor I, no significant format effects were found with either untrained or trained raters. For Instructor II, examination of the analyses involving untrained raters revealed one format difference, showing MSS as more contaminated ($p < .001$, for both MSS coding systems) by halo than BES. The results concerning ratings of Instructor II by trained raters indicated no format differences with respect to halo.

Training effects--For Instructor I, none of the differences between training groups were statistically significant for any scale format. For Instructor II, trained raters using MSS exhibited less ($p < .001$, for both MSS coding systems) halo than untrained raters using MSS, whereas BES and GRS showed no differences between training groups.

All in all, the findings concerning halo do not lend themselves to a general interpretation in that the results for each instructor were inconsistent. A Training x Format interaction was found with Instructor II, but not with Instructor I. Thus, evidence to either support or discredit the hypothesized interaction was inconclusive. With respect to training, although the means indicated that trained raters had generally more variability in their ratings than untrained raters, the differences were not statistically significant for Instructor I and significant with only MSS for Instructor II. One consistent finding that did emerge from the analyses of both instructors was that the formats exhibited no significant differences in halo when the ratings were performed by trained raters.

Interrater Agreement

A series of multivariate and univariate tests of variance homogeneity, along with Wilcoxon tests, were performed to examine the effect of Training and Rating Scale on interrater agreement. As before, the first set of analyses examined form effects, while the second set examined format effects. Both Instructor and Training were held constant for each analysis. A third set of analyses examined training effects, wehrein both Instructor and Rating Scale were held constant for each analysis.

It should be noted that the multivariate test of homogeneity of the within variance/covariance matrices was not performed when there were only five cases in one or both of the groups, as the *N* was too small. Specifically, these groups produced singular matrices of rank 4, as opposed to a full rank of 9 for the nonsingular matrices. The power of the test is greatly reduced in such instances of severe reduction in rank.

Form effects--For both instructors and in both training conditions, analyses concerning differences in interrater agreement between alternate forms revealed no significant findings for either BES or MSS. Since none of the multivariate or univariate analyses revealed significance at the .01 level, the results have not been entered into a table.

Format effects--As mentioned above, in order to maintain a certain independence among the following analyses and avoid confounding format effects with form effects, each form of BES was only compared to the corresponding MSS form. Both forms of BES and MSS were compared to GRS.

The results of tests for format effects are presented in Tables 7 through 10, which are separated according to instructor and training condition. For each comparison in Tables 7 through 10, a rating scale is identified as being either "Scale 1" or "Scale 2", depending on the heading it is under. Each of the alphabetical labels (A through I) corresponds with one of the performance dimensions. A letter appearing under Scale 1 indicates that, for the dimension represented by the letter, the variance in the ratings on Scale 1 was significantly greater than the variance in the ratings on Scale 2, suggesting less interrater agreement among the raters using Scale 1. A letter appearing under Scale 2 is interpreted in the same manner,

Table 7

Significance of Variance Differences (Interrater Agreement) Between Rating Scales

Instructor I (Untrained Raters)

Rating Scales		Multivariate Test of Homogeneity	Dimensions[a] With Variance Differences		Wilcoxon Test
Scale 1	Scale 2		Scale 1[b]	Scale 2[c]	
BSS(A) vs	GRS	.01[d]	B*		ns.
BES(B) vs	GRS	ns.[e]			
BES(A) vs	MSS(A)	.001			ns.
BES(A) vs	revised MSS(A)	.001			ns.
BES(B) vs	MSS(B)	ns.			
BES(B) vs	revised MSS(B)	ns.			
GRS vs	MSS(A)	.01	G*	B**,E*	ns.
GRS vs	revised MSS(A)	.01	G*	B*	ns.
GRS vs	MSS(B)	.01	H*	B**, E**	ns.
GRS vs	revised MSS(B)	ns.			

[a] A=Assignments, B=Attitude Towards Subject , C=Grades, D=Instructor Knowledge, E=Manner of Presentation, F=Objectiveness, G=Organization, H=Student-Teacher Relations, I=Tests

[b] Sample variance was larger for Scale 1 (less interrater agreement)

[c] Sample variance was larger for Scale 2 (less interrater agreement)

[d] Level of statistical significance $(p \leq)$

[e] Not significant at the .01 level

* $p < .02$

** $p < .002$

Table 8

Significance of Variance Differences (Interrater Agreement) Between Rating Scales

Instructor I (Trained Raters)

Rating Scales			Multivariate Test of Homogeneity	Dimensions[a] With Variance Differences		Wilcoxon Test
Scale 1		Scale 2		Scale 1[b]	Scale 2[c]	
BES(A)	vs	GRS	ns.[d]			
BES(B)	vs	GRS	ns.			
BES(A)	vs	MSS(A)	ns.			
BES(A)	vs	revised MSS(A)	ns.			
BES(B)	vs	MSS(B)	ns.			
BES(B)	vs	revised MSS(B)	.001[e]	*G	*F	ns.
GRS	vs	MSS(A)	ns.			
GRS	vs	revised MSS(A)	ns.			
GRS	vs	MSS(B)	ns.			
GRS	vs	revised MSS(B)	.01	*G		ns.

[a] A=Assignments, B=Attitude Toward Subject, C=Grades, D=Instructor Knowledge, E=Manner of Presentation, F=Objectiveness, G=Organization, H=Student-Teacher Relations, I=Tests

[b] Sample variance was larger for Scale 1 (less interrater agreement)

[c] Sample variance was larger for Scale 2 (less interrater agreement)

[d] Not significant at the .01 level

[e] Level of statistical significance ($p \leq$)

* $p < .02$

Table 9

Significance of Variance Differences (Interrater Agreement) Between Rating Scales

Instructor II (Untrained Raters)

Rating Scales		Multivariate Test of Homegeneity	Dimensions[a] With Variance Differences		Wilcoxon Test
Scale 1	Scale 2		Scale 1[b]	Scale 2[c]	
BES(A)	vs GRS	singular[d]	F*		ns.
BES(B)	vs GRS	singular			ns.
BES(A)	vs MSS(A)	singular	D**, E**		p < .01[e]
BES(A)	vs revised MSS(A)	singular	E**		p < .01[e]
BES(B)	vs MSS(B)	singular			ns.
BES(B)	vs revised MSS(B)	singular			ns.
GRS	vs MSS(A)	singular	D**, E**		ns.
GRS	vs revised MSS(A)	singular	E**		ns.
GRS	vs MSS(B)	singular			ns.
GRS	vs revised MSS(B)	singular		F**	ns.

[a] A=Assignments, B=Attitude Towards Subject, C=Grades, D=Instructor Knowledge, E=Manner of Presentation, F=Objectiveness, G=Organization, H=Student-Teacher Relations, I=Tests

[b] Sample variance was larger for Scale 1 (less interrater agreement)

[c] Sample variance was larger for Scale 2 (less interrater agreement)

[d] Multivariate test inappropriate due to singular matrices with greatly reduced ranks

[e] MSS(A) showed better interrater agreement than BES(A)

* $p < .02$

** $p < .002$

Table 10

Significance of Variance Differences (Interrater Agreement) Between Rating Scales

Instructor II (Trained Raters)						
Rating Scales			Multivariate Test of Homogeneity	Dimensions[a] With Variance Differences		Wilcoxon Test
Scale 1		Scale 2		Scale 1[b]	Scale 2[c]	
BES(A)	vs	GRS	.01[d]			ns.
BES(B)	vs	GRS	ns.[e]			
BES(A)	vs	MSS(A)	ns.			
BES(A)	vs	revised MSS(A)	ns.			
BES(B)	vs	MSS(B)	.01		$\overset{*}{G}$	ns.
BES(B)	vs	revised MSS(B)	ns.			
GRS	vs	MSS(A)	ns.			
GRS	vs	revised MSS(A)	.01	$\overset{*}{A}$,$\overset{*}{F}$		ns.
GRS	vs	MSS(B)	.001		$\overset{*}{B}$	ns.
GRS	vs	revised MSS(B)	ns.			

[a] A=Assignments, B=Attitude Towards Subject, C=Grades, D=Instructor Knowledge, E=Manner of Presentation, F=Objectiveness, G=Organization, H=Student-Teacher Relations, I=Tests

[b] Sample variance was larger for Scale 1 (less interrater agreement)

[c] Sample variance was larger for Scale 2 (less interrater agreement)

[d] Level of statistical significance ($p \leq$)

[e] Not significant at the .01 level

* $p < .02$

with interrater agreement being less for Scale 2. (Note that greater variance indicates less interrater agreement). The level of significance for each dmension in indicated directly above the letter.

The findings reported in each table suggest that none of the formats exhibited any general differences in interrater agreement across performance dimensions. In terms of significant variance differences, no scale ever differed from another scale by more than two dimensions. In all of the analyses, in fact, the variance ratio Fs were typically rather small. In many cases, the scale with the larger variance on one dimension emerged as having the smaller variance on a different dimension, as indicated by the nonsignificant Wilcoxon tests. The only significant Wilcoxon test is reported in Table 9, where MSS (A) exhibited better interrater agreement than BES (A). This finding, however, was not replicated across forms nor across instructors, thus leading one to question its reliability.

Tables 7 through 10 were also examined to determine whether or not significant variance differences between particular formats were consistently found for specific dimensions. Such differences did not appear to be tied to specific dimensions across independent estimates of interrater agreement. For GRS vs. MSS (A) in Tables 7 and 9 (untrained raters), for example, although dimension "E" showed significant variance differences across instructors, the differences were not consistent. That is, MSS (A) showed less interrater agreement on dimension "E" with Instructor I, whereas GRS showed less interrater agreement on dimension "E" with Instructor II.

Training effects--The results concerning the effects of training are reported in Table 11. A letter appearing under "Untrained" indicates that, for the dimension represented by the letter, the variance in the ratings by untrained raters was significantly greater than the variance in the ratings by trained raters, suggesting less interrater agreement among untrained raters. A letter appearing under "Trained" is interpreted in a similar fashion, with interrater agreement being less among trained raters. Again, the level of significance for each dimension in indicated directly above the letter.

The results for both instructors suggest that, for each rating scale, training did not have a substantial impact on interrater agreement across the performance dimensions. Even for ratings of Instructor II by students using MSS, where training demonstrated its strongest effect on reducing halo, only two dimensions of MSS (A) revealed significantly greater variance in the ratings by trained raters. The Wilcoxon tests also indicate that there were no statistically significant differences between trained and untrained raters.

Thus, the findings suggest that there were no general form effects, format effects, or training effects with respect to interrater agreement.

Table 11

Significance of Variance Differences (Interrater Agreement) Between Trained and Untrained Raters

Rating Scale	Multivariate Test of Homogeneity	Dimensions[a] With Variance Differences		Wilcoxon Test
	Instructor I			
		Untrained[b]	Trained[c]	
BES(A)	.01[d]			ns.
BES(B)	ns.[e]			
GRS	.01		$\overset{**}{B}$, $\overset{*}{C}$	ns.
MSS(A)	.01		$\overset{*}{G}$	ns.
revised MSS(A)	ns.			
MSS(B)	ns.			
revised MSS(B)	ns.			
	Instructor II			
BES(A)	singular[f]			ns.
BES(B)	singular	$\overset{*}{F}$		ns.
GRS	singular			ns.
MSS(A)	singular		$\overset{**}{D}$, $\overset{**}{E}$	ns.
revised MSS(A)	singular		$\overset{**}{E}$	ns.
MSS(B)	singular			ns.
revised MSS(B)	singular			ns.

[a] A=Assignments, B=Attitude Towards Subject, C=Grades, D=Instructor Knowledge, E=Manner of Presentation, F=Objectiveness, G=Organization, H=Student-Teacher Relations, I=Tests

[b] Sample variance was larger for untrained raters (less agreement)

[c] Sample variance was larger for trained raters (less agreement)

[d] Level of statistical significance ($p \leq$)

[e] Not significant at the .01 level

[f] Multivariate test inappropriate due to singular matrix with greatly reduced rank

* $p < .02$

** $p < .002$

Dimension effects -- Borman (1979) reported "that raters evaluated ratees (vignettes) significantly more accurately on some dimensions than on others, and that for the most part these differences were consistent across formats..." (p. 419). The present study examined the nine dimension variance scores (interrater agreement) of each rating scale to determine whether certain dimensions exhibited consistently better or worse interrater agreement across different rating scales. The Friedman two-way analysis of variance by ranks (Siegel, 1956) was used to test the hypothesis that interrater agreement may be significantly better on some dimensions than on others, even though each rating group used a different rating scale (i.e., BES(A), BES(B), MSS(A), MSS(B), and GRS). The five rating groups represented the N=5 subjects (rows) and the nine dimensions represented the k=9 conditions (columns) for the Friedman test. That is, the variance (interrater agreement) scores of the dimensions were ranked separately for each rating scale.

Eight analyses were performed, two for each training condition of each instructor. One analysis in each case included the original MSS coding system, whereas the other included the revised MSS coding system. None of the eight analyses were significant at $p < .05$, thus suggesting that consensus on an instructor's performance is not consistently better for some dimensions than for others, with either untrained or trained raters. In other words, the relative magnitudes of the dimension variances appeared to vary in a random fashion from one rating scale to the next.

In speculating as to why this occurred, recall that all the dimensions had been previously identified in the scale development phase as areas which students were capable of evaluating. Each dimension, however, still exhibited a considerable number of specific behaviors with rather low rater consensus on perceived level of performance, which were therefore not selected for use in the BES and MSS. Given that students reported they were able to effectively evaluate each dimension and yet unable to adequately agree on the performance level of many of the behaviors exemplifying each dimension, it is conceivable that the instructor's performance on any given dimension was simply not substantially more unequivocal relative to the perceived performance of the other dimensions for significant differences in interrater agreement among the dimensions to emerge. This may not be the case for all instructors, however. If an instructor gives assignments that are strongly disliked by the entire class, for example, a significant dimension effect might well emerge due to consistently high interrater agreement across rating scales on the "assignments" dimension relative to interrater agreement on other dimensions.

Summary

The design of the present study, along with the adopted statistical procedures, provided a unique opportunity to examine the replicability of an effect over dimensions, alternate forms, and ratees (instructors). With respect to replication over dimensions, the multivariate analyses of leniency/severity and interrater agreement were used as initial tests of group differences. When a multivariate difference proved to be nonsignificant, further investigations were not pursued. On the other hand, when a significant multivariate difference was found, subsequent univariate and nonparametric tests were performed to disclose the nature of the difference, as the interest was not in a multivariate difference per se, but rather in a particular kind of multivariate difference.

At the multivariate level of analysis, the results showed several significant format and training effects, which have also been found in other studies focusing on the "big picture" by combining the data of different dimensions as well as different ratees into some kind of single, aggregate analysis. In this study, however, subsequent examinations revealed that, in most cases, significant group differences at the univariate level were specific to one or two dimensions only, and that the particular dimensions showing statistical significance differed from one analysis to another, thus indicating that group differences were not consistently showing up on a specific subset of dimensions. The nonparametric tests, which were used to statistically determine the pervasiveness or generalizability of an effect over all dimensions irrespective of the statistical significance of the effect on each separate dimension, further indicated that, in most cases, the effect did not exhibit an acceptable level ($p \leq .01$) of consistency in the direction of differences across the performance dimensions. Thus, even though statistically significant differences may have been present between two rater groups, the findings were interpreted as showing no form, format, or training differences regarding leniency/severity and interrater agreement when only one or two dimensions showed a significant difference and the direction of the differences over all dimensions did not attain statistical significance in the nonparametric test.

The basic logic for interpreting significance based on replication over dimensions was also applied when examining the replicability of an effect over forms and instructors. Replication over forms and instructors would increase our confidence that the effect was due to the treatment rather than to other extraneous variables, such as a priori differences in rater groups (even though raters were randomly assigned to each group) or peculiar characteristics of a particular instructor.

The research design in this study (i.e., many raters rating one ratee) is unable to effectively separate the effects of an experimental manipulation from the possibly uncommon characteristics of a particular ratee (Saal et al., Note 2). It therefore becomes important to include more than one ratee, as the emergence of generally strong halo (variance across the dimensions in each rater's ratings) across all formats and training conditions could conceivably result from the actual characteristics of the ratee being evaluated. The overall impact of any peculiar characteristics of a ratee on the general findings is diminished as the number of ratees is increased. It would be even more desirable to have the raters rate more than one ratee, however, as a rater group's tendency to repeat a bias over several ratees could then be assessed.

The effects of forms, formats, and training are summarized in Table 12. The results concerning leniency/severity and interrater agreement indicated that the significant findings failed to demonstrate replicability either over alternate forms or over instructors, therefore suggesting that the significant findings may be quite unreliable and, therefore, of questionable "practical" significance. The results concerning halo indicated that although a Format x Training interaction was found with one instructor, there was no Format x Training interaction with the other instructor, thus demonstrating no clear support either for or against the interaction hypothesis. As mentioned before, however, the results for both instructors suggest that there were no format differences when trained raters performed the ratings.

Table 12

Effects of Rating Scale and Training:

A Summary

	Instructor I	Instructor II
Alternate Forms		
Leniency/Severity:		
Untrained Raters	No differences	No differences
Trained Raters	No differences	No differences
Halo:		
Untrained Raters	No differences	No differences
Trained Raters	No differences	No differences
Interrater Agreement:		
Untrained Raters	No differences	No differences
Trained Raters	No differences	No differences
Scale Formats		
Leniency/Severity:		
Untrained Raters	BES(A) $>$ [a]MSS(A) GRS $>$ [a]MSS(A)	No differences
Trained Raters	No differences	No differences
Halo:		
Untrained Raters	No differences	BES $<$ [b]MSS
Trained Raters	No differences	No differences
Interrater Agreement:		
Untrained Raters	No differences	MSS(A) $>$ [c]BES(A)
Trained Raters	No differences	No differences

Table 12 (cont.)

	Instructor I	Instructor II
Training		
Leniency/Severity:		
BES	No differences	No differences
GRS	No differences	No differences
MSS	No differences	No differences
Halo:		
BES	No differences	No differences
GRS	No differences	No differences
MSS	No differences	Trained <[b]Untrained
Interrater Agreement:		
BES	No differences	No differences
GRS	No differences	No differences
MSS	No differences	No differences

Note: The results found with the original MSS coding system were consistent with the results found with the revised MSS coding system.

[a] > = showed higher mean ratings than

[b] < = showed less halo error than

[c] > = showed greater interrater agreement than

Discussion

The issues addressed in the present study can be separated into three general areas: form effects; format effects; and training effects. The findings of this study have been subdivided according to these three areas in the summary section for easy reference.

Form effects

Very little research has examined the impact of the choice of particular behavioral statements on performance ratings (Zedeck et al., 1976). The question concerns whether forms with qualitatively different behavioral statements produce similar ratings. According to the scale development procedures and the adopted selection criteria for behavioral statements, each surviving statement should adequately exemplify performance at a specified level for a specified dimension. Oftentimes, however, several statements qualify for selection at a given performance level, and the choice becomes somewhat arbitrary.

The issue is particularly germane for MSS, as no dimension labels, definitions, or graphic scales are provided, and each dimension is evaluated on the basis of three behavioral statements alone. For MSS, corresponding behavioral statements of the two forms, though describing different behaviors, are assumed to exemplify essentially the same level of performance for the same dimension. It is obvious from both a theoretical and practical point of view that the adopted procedure for developing behaviorally based rating scales would be of limited value if the ratees' evaluations were affected by the selection of a particular behavioral description from a set of descriptions which all purportedly represented the same level of performance on the same dimension.

Suppose, for example, that form differences were found for a particular dimension. Aside from random variation, it is possible that the behavioral statements of each form were measuring either different performance levels or different dimensions. This would suggest that either the selection criteria for the behavioral statements did not effectively group behaviors exemplifying the same dimension and performance level or the identified performance dimension was too general (the dimension actually contained at least two behaviorally distinct dimensions). This would necessitate a reevaluation of the scale development procedures and may invalidate these procedures.

As in Zedeck et al. (1976), the results of this study indicated that the adopted scale development procedures were able to produce alternate forms of behaviorally based rating scale formats which evinced comparable psychometric properties. Specifically, the alternate forms of a particular format yielded ratings with comparable levels of leniency, halo, and interrater agreement. Although the analyses performed were far from a complete test of parallelism (Ghiselli, 1964) and used the questionable practice of testing for the null hypothesis, it is still informative in a practical sense to obtain results which suggest that the different forms provided comparable ratings.

The results concerning the effects of training on alternate rating forms lend further support to their comparability in that the forms of a particular format appeared to be similarly affected by an experimental manipulation. The analyses revealed the following:

First, for each format, the difference between trained and untrained raters who used Form A was consistent with the difference between trained and untrained raters who used Form B. That is, if a significant overall difference was found with one form, it was also found with the other form, as demonstrated in the significant training effect for both forms of MSS in the halo analyses involving Instructor II.

Secondly, as with untrained raters, no form differences were found in the ratings by trained raters for either BES or MSS.

Since these findings suggest that any appropriate subset of the behavioral descriptions satisfying the selection criteria could be selected without fear that the resultant ratings would be influenced, one might therefore consider developing alternate forms of the behaviorally based rating scales, thereby utilizing more of the behavioral descriptions. One advantage is that the raters would not be rating the ratees on the same behaviors over and over again, which may help the raters to maintain a higher level of attention to each behavioral example while performing the evaluations.

A second advantage is that the ratees would be more knowledgeable of the specific kinds of behaviors that raters generally consider as more or less desirable, which may better help them to improve their performance.

Future research should examine the effects of different behavioral statements for MSS more directly and at a more individual level. This could be done by having each rater use both forms of MSS to evaluate the ratees. Such research could determine, for each dimension, whether or not individual raters perceive the qualitatively different behavioral descriptions of each form as measuring the same general area of performance and as measuring comparable levels of performance. Such research is essential to determining the feasibility of evaluating a performance dimension on the basis of three specific behavioral descriptions alone.

Although such research is essential for BES as well, the appropriateness of a within-groups design is somewhat questionable, as the accompanying graphic scale provides a reference for the raters to maintain rating consistency from one form to another, regardless of the behavioral descriptions. That is, ratings on the second form are subject to contamination from knowledge of where the rating was located graphically on the first form. Perhaps future research could devise a within-groups design for BES wherein such potential contamination is minimized.

Format effects

Previous research has been unable to clearly identify a rating scale format with relatively superior psychometric properties (Borman and Vallon, 1974; Burnaska and Hollmann, 1974; Borman and Dunnette, 1975; Bernardin, 1977; DeCotiis, 1977; Finley et al., 1977; Saal, 1979; and Borman, 1979). The results of this study also reveal no clear support for the psychometric superiority of a particular rating scale format, as a format effect failed to demonstrate replicability across forms and/or instructors.

For leniency/severity, although the ratings of Instructor I by untrained raters were generally more lenient on BES(A) and GRS than on MSS(A), the findings were not replicated across either forms or instructors, thus leading one to question the practical significance and even the genuineness of the effect. The results concerning halo were also mixed in that a format effect was only found in the ratings of Instructor II by untrained raters. Furthermore, the significant finding indicated that BES exhibited less halo than MSS, which is inconsistent with previous results showing BES as more contaminated by halo (Saal and Landy, 1977; and Saal, Note 4). Regarding interrater agreement, only one or two dimensions showed significant differences in each comparison, and only two Wilcoxon tests were statistically significant, wherein untrained raters using MSS(A) exhibited greater interrater agreement than untrained raters using BES(A) for both MSS coding systems. Based on the nonreplicability of the significant Wilcoxon tests over both forms and instructors, however, the author is inclined to conclude that the significant findings were either highly specific or spurious.

Although the results pertaining to untrained raters were not consistent over forms and/or instructors, the results concerning trained raters consistently showed no significant format effects across levels of both forms and instructors.

The nonsignificant or inconsistent findings of this study and the inconclusive results of previous studies indicate that of the formats usually reported in the literature thus far (e.g., BES, GRS, MSS, and Summated Rating Scales), no single format can be generally regarded as having demonstrated clear superiority in relative psychometric quality, thus suggesting that a particular format does not appear to be the major determinant of better performance ratings.

Part of the reason for the inconsistent findings among the different studies, however, may be due to differences in the operational definitions of halo, leniency/severity, central tendency, restriction of range, and interrater agreement, as a review of related studies (Saal et al., Note 2) revealed that several methods of measurement have been used to operationalize each construct. Although the disparate operational definitions may have resulted from conceptual differences, they were also an unavoidable consequence of differences in the experimental designs of each study. Researchers are oftentimes forced to adopt partial designs that involve some variation of a completely crossed Rater x

Ratee x Dimension matrix, which usually suffer from an extremely small number of raters or ratees, thus precluding the researchers from performing certain types of analyses. Saal et al. (Note 2) have demonstrated that different techniques for quantifying a particular construct can lead to inconsistent conclusions. This finding therefore suggests that comparisons between the results of different studies are less meaningful, and possibly meaningless, when the studies use different quantification procedures for the same construct.

The present study does not allow such analyses as dimension intercorrelations with ratees as data points and ANOVAS examining Ratee main effects and interactions. In fact, it is precluded from any type of analysis which requires a distribution of ratees. Thus, this design may have limited value for comparisons with studies where comparable analyses could not be performed.

Another problem related to the quantification problem concerns statistical methodology. It is important that the statistical procedures used to examine the data be carefully chosen, as inappropriately used statistical tests can also lead to erroneous conclusions. Careful consideration should therefore be given to the nature of the data and the kinds of assumptions that can be made before proceeding to select the statistical tests. At least one can then feel confident that the questions asked of the data were answered by an appropriate statistical model.

Training effects

Previous research has generally shown that training can reduce halo and leniency without necessarily affecting interrater agreement or rating accuracy (Bernardin, 1978; Bernardin and Walter, 1977; Borman, 1975; Borman, 1979; and Latham et al., 1975). The results of this study indicated that a brief training session did not significantly affect interrater agreement, nor did it affect leniency/ severity. With respect to halo, while the means were generally in the predicted direction, a significant training effect was only found with raters using MSS to evaluate Instructor II; the effect was not replicated with Instructor I. Thus, it appears that brief rater training produced a rather weak and inconsistent effect in reducing halo.

A study by Borman (1979) found that even an intensified training program was not consistently effective in reducing halo across two rating tasks. Borman's (1979) results concerning halo further indicated at a more specific level that while training might be quite effective with a particular format in one case, it may be completely ineffective with that format in another, which is consistent with the findings of this study.

Based on the results of this study as well as other studies (Borman, 1979 and Vance et al., 1978), it appears that trained raters do not consistently produce ratings that are superior in psychometric quality to the ratings of untrained raters. It should again be noted, however, that differences among the studies in the operational definitions and statistical methodologies may possibly be partly responsible for the inconsistent findings.

Conclusions

Studies examining rating scale format and/or rater training have not been very successful at accounting for differences in rating quality. The problem in this study was not that there were no differences, as revealed by an inspection of the halo scores of individual raters, rather it was that within each rater group, individual raters exhibited sizable differences in halo. Some raters within each group exhibited relatively strong halo, whereas others in each group exhibited minimal halo.

The obvious implication is that since the within-treatments variance appears to be as substantial as the between-treatments variance, more attention should therefore be focused on the raters themselves. That is, the individual characteristics of raters should be viewed as a variable worthy of investigation, not as an error term only. Cronbach's (1957) admonishments were undoubtedly directed as much toward the area of performance appraisal as toward any other field of scientific psychology. It is becoming apparent that the goal of explaining differences in rating quality as well as improving rating quality is not about to be realized by conveniently covering up individual differences in the error term and quietly referring to such differences (though very appopriately) as unexplained variance. A factorial experiment can hardly reveal anything worthwhile if a large part of the answer has been cast into the error term, regardless of the statistical legitamacy of the maneuver.

Very few studies have treated rater characteristics as something other than just an error term (Bernardin, 1978 and Schneier, 1977). Specifically, more research should be directed toward examining rating quality as a function of such rater characteristics as cognitive complexity, self-confidence in ability to evaluate performance, perceived importance of evaluations, etc. Moreover, as in Schneier (1977), the effects of rating scale format could be examined in such studies as well, with the goal being to identify possible Format x Rater Characteristic interactions. Such an interaction may provide a possible explanation for the inconsistent findings among the studies examining format effects. One might contend, as other researchers have (Zedeck et al., 1976), that time and effort are not well spent if the focus is on format differences alone, as too much of the answer concerning rating quality appears to lie elsewhere.

Future research could also investigate the possible impact of such antecedent rater characteristics as cognitive complexity on the effectiveness of training raters to improve their rating skills. Inclusion of different training techniques (Bernardin and Walter, 1977; Borman, 1979; and

Latham et al., 1975) might well reveal a Training x Cognitive Complexity interaction. For example, given that a cognitively complex rater is by definition more able to dimensionalize ratee performance, and that training can effectively improve a rater's ability to dimensionalize ratee performance, such research may show that while an intensified training program may prove to be significantly more effective than a brief training program for cognitively simple raters, the effects of a brief and intensified training program may be comparable for cognitively complex raters. Moreover, the psychometric quality of the ratings by cognitively complex raters receiving training may not differ substantially from the rating quality of cognitively complex raters receiving no training, as the untrained raters may already be effectively dimensionalizing performance. The rationale behind the hypothesized differential effect of training techniques for cognitively simple raters is that an intensified training program may be required to fully induce such raters to effectively observe and evaluate performance in a multidimensional manner.

One might also consider including a Rating Scale Format factor along with the Training and Cognitive Complexity factors, as such a design would be even more informative and revealing.

Clearly, the effectiveness of rating scale format and rater training in improving the psychometric quality of ratings is indeed mediated by rater characteristics and should therefore be examined as a function thereof.

The practical value of such research is that the direction is towards attempting to identify which of the treatments or treatment combinations (i.e., training and format) is most effective and efficient for each type of raters so that the rating effectiveness of all raters might be maximized, as opposed to searching for the most effective and efficient treatments for all raters in general and thereby possibly allowing the rating effectiveness of many raters to remain substandard.

As a final caution, the context within which a study was conducted should not be lightly dismissed. The findings of this study, which were based on student ratings of instructors, may not generalize to the typical supervisor-subordinate rating situation. Warmke and Billings (1979) found that even within the same organizational context, findings based on experimental ratings were not consistent with findings emerging from administrative ratings. Thus, one must be discreet in generalizing to less comparable contexts, as the context is unquestionably of great importance in determining the "real" effects of experimental manipulations.

Appendix I

Procedures for developing Behavioral Expectation Scales (BES)

1. Performance dimensions. A "representative" sample of the rater population is gathered together to identify (in one or two words) each important characteristic or aspect of the job which they feel should be considered when appraising performance. The participants then generate behaviorally oriented definitions for each dimension label, and redundant dimensions or dimensions that fail to achieve consensual agreement on a definition are eliminated.

2. Behavioral statements. A second group is instructed to submit behavioral statements for each dimension which exemplify various levels of performance, i.e., superior, moderate, and poor. The statements are then edited by those involved in constructing the scales into the form of behavioral expectations. That is, instead of such statements as, "The employee admits mistakes when he/she makes them.", they would be rewritten as, "This employee can be expected to admit mistakes when he/she makes them." Editing the statements into expectations at this point is optional, however, and can be postponed until a later step, or the editing can be eliminated altogether, in which case the behaviors are used in their original form (Dickinson and Tice, 1973; Landy, Farr, Saal, and Freytag, 1976). Scales that do not contain "expectation" statements are identified as behaviorally anchored rating scales (BARS), rather than behavior expectation scales (BES). Also, a number of studies (Borman and Dunnette, 1975; Borman and Vallon, 1974; Campbell, Dunnette, Arvey

and Hellervik, 1973; DeCotiis, 1977; Fogli, Hulin, and Blood, 1971; Keaveny and McGann, 1975; Motowidlo and Borman, 1977) eliminated step 1 so that participants would attend more to specific behaviors rather than general aspects of a particular activity. In this case, raters are initially asked to generate behavioral descriptions of various levels of job performance. The raters and/or project directors then group the behavioral statements into categories and provide the labels and definitions.

3. Retranslation of statements. A third group from the rater population is presented with a list of the defined dimensions and a single randomized list of the behavioral expectation statements from all the dimensions and instructed to indicate the dimension to which each statement "belongs". The retranslating (or reallocating) of the behavioral statements is performed individually rather than collectively by the entire group. A retranslation criterion is set (usually 50% - 80%), and statements are eliminated if they do not satisfy the adopted criterion of reassignment. A dimension is subsequently eliminated if an insufficient number of statements are successfully reallocated to it. That is, a dimension is subject to elimination if it is determined that there are an insufficient number of behavioral statements to effectively "anchor" the dimension's performance continuum.

4. Scaling of statements. The surviving dimensions, along

with their definitions and retranslated statements, are presented to another group. Each person in the group numerically rates (usually on a 7 or 9- point scale) each statement according to the level of performance exemplified by the behavior in relation to the dimension it describes. The mean rating of each statement is taken to represent the level of performance indicated, while the standard deviation is used to describe the extent to which raters agree/disagree upon that level. A lower standard deviation suggests greater agreement. A standard deviation criterion is set (usually 1.5 to 2.0), and statements are again eliminated if their standard deviations do not satisfy this criterion. Standard deviations greater than the set criterion suggest unacceptable amounts of variation and imply too much disagreement regarding the level of desirability. If an excessive amount of variability is generally the case for nearly all of the statements representing a particular dimension, then the dimension is again subject to elimination.

5. Anchoring dimension scales. Each remaining dimension is typically accompanied by a vertically positioned graphic-type rating scale. The scale is usually anchored with both numbers and equally spaced behavioral statements, if possible. The mean rating of the statement determines its location on the scale. The instrument in final form consists of a series of the defined dimensions and their corresponding scales.

Appendix II

Questionnaire for Gathering Scaling Data on Behavioral Statements

SCALING OF BEHAVIORAL STATEMENTS

On the following sheets are shown behavioral statements which represent various levels of performance. These statements are grouped together under nine (9) dimensions concerned with teaching effectiveness. Please read the definition of the first dimension, go to the behavioral statements for that dimension and rate each statement on a scale of 1 to 7 (7=superior; 4=average; 1=poor) as to the level of performance it represents. Then go through each of the remaining dimensions doing the same thing.

1. ASSIGNMENTS: Extent to which the professor is clear on what is to be done, avoids assigning excessive amounts, and provides assignments which contribute to the understanding of the subject matter rather than just providing busy work.

2. ATTITUDE TOWARDS SUBJECT: Extent to which the professor shows personal interest in the material and displays a positive attitude towards teaching the subject.

3. GRADES: Extent to which the professor's grading practices remain consistent and free of confusion and are also fair.

4. INSTRUCTOR KNOWLEDGE: Extent to which the professor is aware of current material related to the course or to his/her field and is able to accurately answer or direct the student to specific sources that will answer questions concerning the subject matter.

5. MANNER OF PRESENTATION: Extent to which the professor's methods of presentation and use of audio-visual aids help emphasize and clarify important points; ability to present material clearly and concisely on a level sudents can understand.

6. OBJECTIVENESS: Extent to which the professor remains objective and presents a fair treatment of all points of view on controversial or debatable topics.

7. ORGANIZATION: Extent to which the professor arranges the subject matter and course objectives in an orderly and logical sequence for thorough coverage.

8. STUDENT-TEACHER RELATIONS: Extent to which the professor shows a true, sincere concern for the welfare of the students through such things as dependability, availability for help, and consideration of student feelings; establishing rapport with the students.

9. TESTS: Extent to which the professor writes clear, unambiguous questions that relate to and are representative of in-class material and outside readings which were stressed adequately in class.

ASSIGNMENTS

____ 1. This instructor can be expected to vaguely give the student an idea of what is to be done on the next assignment and not really say when it is due.

____ 2. This instructor can be expected to give an assignment and then discuss it with the students as soon as possible.

____ 3. This instructor can be expected to assign a lot of problems that seem to cover the same thing.

____ 4. This instructor can be expected to assign homework which pertains only to materials that will be tested over.

____ 5. This instructor can be expected to give assignments that offer at least some understanding of the subject matter.

____ 6. This instructor can be expected to clearly explain what is to be done and assign enough work to understand the concepts to be tested on and needed later.

____ 7. This instructor can be expected to give assignments that cover material much deeper than the students need to learn.

____ 8. This instructor can be expected to give assignments so that the students know exactly what is expected of them.

____ 9. This instructor can be expected to assign problems and then not return them or discuss any of them.

____ 10. This instructor can be expected to give assignments that have personal value to the students apart from the subject matter itself.

____ 11. This instructor can be expected to tell students which problems in the assignment are more important.

____ 12. This instructor can be expected to assign excessive amounts of homework.

____ 13. This instructor can be expected to give assignments that relate to the subject matter and also provide further understanding.

____ 14. This instructor can be expected to assign very difficult homework.

____ 15. This instructor can be expected to fluctuate between too much and too little in assignments.

____ 16. This instructor can be expected to make interesting assignments.

____ 17. This instructor can be expected to give only busy work.

____18. This instructor can be expected to give plenty of time for students to get assignments done.

____19. This instructor can be expected to assign questions that do not have to be turned in, but still discusses them in class.

____20. This instructor can be expected to clearly explain the assignment at the end of class.

____21. This instructor can be expected to assign too much work once in a while.

ATTITUDE TOWARDS SUBJECT

____ 1. This instructor can be expected to show dislike for the subject and teaching and act as though he/she got stuck with it.

____ 2. This instructor can be expected to show a "take it or leave it" attitude and not show any enthusiasm.

____ 3. This instructor can be expected to be really excited about what is taught.

____ 4. This instructor can be expected to be enthused about his/her own area of study but is straight forward on the rest of the material.

____ 5. This instructor can be expected to show no interest and act like he/she does not care for the subject at all.

____ 6. This instructor can be expected to tell students that he/she enjoys the work.

____ 7. This instructor can be expected to show interest in the subject but seems to lack any great involvement.

____ 8. This instructor can be expected to be really into what he/she is doing and seem to enjoy it.

____ 9. This instructor can be expected to look almost too tired to even be able to lecture on the material.

____10. This instructor can be expected to, for the most part, enjoy the subject but may show a negative attitude towards it once in a while.

____11. This instructor can be expected to act as if he/she has much better things to be doing.

____12. This instructor can be expected to always show an interest in what he/she is saying and doing.

____13. This instructor can be expected to be late for class many times.

____14. This instructor can be expected to say "this is going to be boring so bear with me".

____15. This instructor can be expected to get excited about the subject matter and personally feel that everyone can benefit from it.

____16. This instructor can be expected to come each day and present the material seeming to enjoy doing it, yet look anxious to leave upon dismissal of class.

____17. This instructor can be expected to at least not look negative towards teaching the subject.

____18. This instructor can be expected to have an attitude like "who cares" when teaching the subject.

____19. This instructor can be expected to show that he/she enjoys the subject and job and the students can tell.

GRADES

____ 1. This instructor can be expected to grade in a way that does not favor any one particular kind of student.

____ 2. This instructor can be expected to use test scores only for determining grades.

____ 3. This instructor can be expected to not let students know how grades are arrived at.

____ 4. This instructor can be expected to base much of the grade on personal opinion.

____ 5. This instructor can be expected to grade according to participation in class, attendance, and test scores.

____ 6. This instructor can be expected to grade strictly on a percentage of total possible (90 - 80 - 70 - 60).

____ 7. This instructor can be expected to explain his/her grading method so the students know how grades are determined and also has evidence for the grades.

____ 8. This instructor can be expected to curve the grades according to the difficulty of the test and the overall student scores.

____ 9. This instructor can be expected to grade with a "C" as average.

____10. This instructor can be expected to not let the students know exactly where they stand throughout the course in terms of a grade.

____11. This instructor can be expected to be more interested in just getting a grade on paper than in how it was arrived at.

____12. This instructor can be expected to give grades based on test scores and homework.

____13. This instructor can be expected to grade with such diversity that questions are always asked at the time of grades.

____14. This instructor can be expected to grade in a way that favors a particular kind of student.

____15. This instructor can be expected to give partial credit for answers that are partially correct.

____16. This instructor can be expected to be a hard grader that requires superior effort on the part of the student for a good grade.

____17. This instructor can be expected to use the same grading scale all the time.

____18. This instructor can be expected to grade more on class participation than on test scores.

____19. This instructor can be expected to use the test scores of the class to figure out grades and not go by total possible.

____20. This instructor can be expected to grade the student more favorably if he/she is trying and shows considerable effort.

____21. This instructor can be expected to maintain high standards even when the grades of the entire class are low in relation to that standard.

____22. This instructor can be expected to change his/her grading practices if most of the students feel it is unfair.

INSTRUCTOR KNOWLEDGE

____ 1. This instructor can be expected to answer questions but rely on rather old material.

____ 2. This instructor can be expected to be up to date in the field and able to relate new concepts and ideas to the text and answer questions concerning them.

____ 3. This instructor can be expected to have read quite a few books on the subject and talked to people who are more knowledgeable.

____ 4. This instructor can be expected to lecture off the same notes used for the last five years which were outdated five years previous to that.

____ 5. This instructor can be expected to have to continually refer back to notes to answer questions.

____ 6. This instructor can be expected to be aware of current material in the area and relate it to previous studies.

____ 7. This instructor can be expected to find answers to most of the students' questions and the answers are fairly accurate.

____ 8. This instructor can be expected to give sudents a good idea of recent developments over the past few years, but is not up to date on present studies.

____ 9. This instructor can be expected to seem not to know what he/she is talking about and can't answer questions.

____10. This instructor can be expected to be aware of new concepts and although unable to answer specific questions, can direct the student to sources that can.

____11. This instructor can be expected to add current material to his/her class presentations which are used as new examples.

____12. This instructor can be expected to know of virtually all the major research projects currently being carried out in his/her field.

____13. This instructor can be expected to know very little of current research efforts in his/her field.

____14. This instructor can be expected to beat around the bush when a student asks a question and will not actually say that he/she does not know the answer.

____15. This instructor can be expected to give the student some idea of where to go to find an answer to a question.

____16. This instructor can be expected to admit not knowing an answer to a question and do his/her best to give the student a source of information.

____17. This instructor can be expected to know about what he/she was lecturing on and not need notes to read off of.

____18. This instructor can be expected to be totally unaware of current material related to his/her field of study.

____19. This instructor can be expected to try and find the answer to a question that he/she does not know.

MANNER OF PRESENTATION

____ 1. This instructor can be expected to use movies, graphs, and charts to help get the ideas across.

____ 2. This instructor can be expected to rarely, if ever, use audio-visual aids.

____ 3. This instructor can be expected to use an overhead projector without explaining what is shown and merely read the information already printed on the overhead.

____ 4. This instructor can be expected to have guest speakers every once in a while.

____ 5. This instructor can be expected to have actual demonstrations in class so the students can experience what they are studying and also has films to illustrate points.

____ 6. This instructor can be expected to go through the material very fast and allow no time for questions.

____ 7. This instructor can be expected to use ausio-visual aids but not clearly explain how it pertains to the lecture.

____ 8. This instructor can be expected to supplement lectures with overhead pictures and graphs and explain them so students are able to understand what is presented.

____ 9. This instructor can be expected to leave the lecture material only occasionally to give an interesting sideview.

____10. This instructor can be expected to go through a detailed analysis of difficult problems and draw pictures to help clarify the concepts.

____11. This instructor can be expected to show films; however, they are usually very old.

____12. This instructor can be expected to not convey the material on a level students can understand.

____13. This instructor can be expected to use aids and examples that relate to the students' time and lifestyle.

____14. This instructor can be expected to show films related to the subject matter but not discuss them afterwards.

____15. This instructor can be expected to put main points and ideas on an overhead projector and then talk about each one.

____16. This instructor can be expected to not give examples to simplify the information given.

____17. This instructor can be expected to use interesting films and discuss them afterwards in order to clarify major points.

____18. This instructor can be expected to just lecture and hardly even use the blackboard.

____19. This instructor can be expected to present the material in a way that is kind of difficult to follow.

____20. This instructor can be expected to do experiments to get a point across.

____21. This instructor can be expected to lecture mostly and show a film every once in a while.

OBJECTIVENESS

____ 1. This instructor can be expected to act as if his/her views are the only correct ones and not listen to other views.

____ 2. This instructor can be expected to always state his/her point of view, although he/she will allow discussion of other points of view.

____ 3. This instructor can be expected to openly discuss each side of the debate and not let personal feelings interfere, but still offer his/her knowledge readily.

____ 4. This instructor can be expected to listen to all points of a discussion, letting each have their own say.

____ 5. This instructor can be expected to talk only about his/her belief on the issue and does not leave any room for debate and discussion.

____ 6. This instructor can be expected to discuss both sides but lean towards his/her beliefs and try to convince others of them.

____ 7. This instructor can be expected to show the various points of view, allow discussion, not put down anyone, and listen and let students decide for themselves.

____ 8. This instructor can be expected to state his/her opinion and openly criticize students having a different one.

____ 9. This instructor can be expected to inform students of conflicting ideas on a topic, but only present a case as to why he/she believes as he/she does.

____10. This instructor can be expected to clearly show favor for a certain side, but still present the other side.

____11. This instructor can be expected to keep the students unaware of his/her opinion on a debatable or controversial topic until after they have stated theirs.

____12. This instructor can be expected to express his/her opinion and encourage students to express their own opinions.

____13. This instructor can be expected to present both sides of the issue and then ignore the side he/she does not favor.

____14. This instructor can be expected to talk about the good and bad points of all topics.

____15. This instructor can be expected to be biased and, although saying he/she welcomes debate, becomes very defensive when questions are raised.

____16. This instructor can be expected to listen to the students' points of view and also make his/her own point of view known.

____17. This instructor can be expected to never let his/her opinion on a debatable or controversial topic be known to the students.

____18. This instructor can be expected to be completely close minded and students that openly disagree with him/her have a very difficult time in the class.

____19. This instructor can be expected to present both sides of a debate in equal detail and without prejudice even if leaning more towards one particular view.

____20. This instructor can be expected to listen to the student's point of view and then argue against it.

ORGANIZATION

____ 1. This instructor can be expected to arrange the subject matter so that the most important material is sure to be covered.

____ 2. This instructor can be expected to hand out a schedule for the semester which shows objectives for tests as well as outside material to be covered in lectures.

____ 3. This instructor can be expected to get confused a little, but on the whole has what he/she is going to say fairly well planned.

____ 4. This instructor can be expected to jump all around and students find it very difficult to follow.

____ 5. This instructor can be expected to have an outlined, detailed schedule of each days' activities to present to the students at the beginning, yet changeable if needed.

____ 6. This instructor can be expected to give a vague sheet of chapter order and then rarely cover the material on schedule or give tests on projected days.

____ 7. This instructor can be expected to have some organization but, at times, get off the topic and discuss something totally unrelated.

____ 8. This instructor can be expected to present material that runs right into the next without skipping all around.

____ 9. This instructor can be expected to skip over necessary information that is needed for concepts later studied.

____10. This instructor can be expected to organize the material into increasing levels of difficulty.

____11. This instructor can be expected to jump around from one thing to the next and frequently get off the subject.

____12. This instructor can be expected to have everything under control in class and tell students at the beginning of class what will be done that day.

____13. This instructor can be expected to have a lecture carefully planned so that it can last the full class period.

____14. This instructor can be expected to present an outline of the course at the beginning of the semester and abide by it without unnecessary postponements or delays.

____15. This instructor can be expected to start talking about the subject without well planned direction and ends up repeating things and getting a little lost.

____16. This instructor can be expected to give an outline for reading assignments and lecture material to be covered for the semester.

____17. This instructor can be expected to present material that sometimes overlaps with materials that has already been presented.

____18. This instructor can be expected to verbally present and explain a general course outline on the first day of class.

____19. This instructor can be exnected to jump back and forth from one chapter to another in the book.

____20. This instructor can be expected to hand out a course outline sheet showing dates when chapters will be covered, test dates, and then sticking to the sheet.

STUDENT-TEACHER RELATIONS

____ 1. This instructor can be expected to not even care if the students do not understand the material.

____ 2. This instructor can be expected to help the students with their problems and listen to their opinions.

____ 3. This instructor can be expected to help students only if they first ask for it.

____ 4. This instructor can be expected to act as if students were inferior instead of having a person to person attitude.

____ 5. This instructor can be expected to know students by their student numbers only and not care to know their names.

____ 6. This instructor can be expected to recognize students who are in his/her classes.

____ 7. This instructor can be expected to care about how students are doing in the class and help them with their questions.

____ 8. This instructor can be expected to remind students that his/her office is open whenever they need to talk.

____ 9. This instructor can be expected to have help sessions and invite students to see him/her individually if they are having trouble with the material.

____10. This instructor can be expected to have the students do what they can and not really help them when they need it.

____11. This instructor can be expected to recognize a student's extra effort in the course.

____12. This instructor can be expected to meet with each student once every two or three weeks and discuss the student's progress.

____13. This instructor can be expected to act like the student is taking up his/her precious time when the student asks for help.

____14. This instructor can be expected to not allow enough time to help students.

____15. This instructor can be expected to have a student come to his/her office after class or wait until the next class for an answer to a specific question.

____16. This instructor can be expected to help the student with any school related problem.

____17. This instructor can be expected to hardly ever be around for help.

____18. This instructor can be expected to treat the student as an equal and always be open for counseling and opinions.

____19. This instructor can be expected to invite students to come and talk with him/her after class or during office hours whenever they have a problem with the class.

____20. This instructor can be expected to be available during his/her office hours.

____21. This instructor can be expected to leave it all up to the students and not be willing to help them.

____22. This instructor can be expected to be aware of the common problems students may have concerning the subject matter and is willing to help the students with them.

____23. This instructor can be expected to be uninterested in a student's achievements.

____24. This instructor can be expected to help students acquire better study habits.

____25. This instructor can be expected to hold grudges against students.

____26. This instructor can be expected to never have help sessions.

____27. This instructor can be expected to take time to talk to each student individually.

____28. This instructor can be expected to be totally ignorant of student problems.

TESTS

____ 1. This instructor can be expected to test only on information given in class.

____ 2. This instructor can be expected to give multiple-choice questions designed to trick students, not to test knowledge.

____ 3. This instructor can be expected to have tests that cover class material with the more important topics covered more often.

____ 4. This instructor can be expected to thoroughly evaluate test questions for difficulty and clarity.

____ 5. This instructor can be expected to write questions that can be answered in two ways, depending on how the student happens to interpret the question.

____ 6. This instructor can be expected to test on too much material at one time.

____ 7. This instructor can be expected to not try to trick students with test questions.

____ 8. This instructor can be expected to test the class on relevant material that has been discussed in class.

____ 9. This instructor can be expected to test too much on picky details instead of on the main ideas of the course.

____10. This instructor can be expected to test on material not covered in class.

____11. This instructor can be expected to allow the students ample time to answer the questions carefully.

____12. This instructor can be expected to give test questions that are very difficult to understand.

____13. This instructor can be expected to test in order to see if students have read the text and kept up with the assignments.

____14. This instructor can be expected to test more heavily on some parts of the material than on others without telling the student.

____15. This instructor can be expected to write clear questions.

____16. This instructor can be expected to throw out a badly worded question.

____17. This instructor can be expected to test on the students' understanding of the major concepts.

____18. This instructor can be expected to give tests that are too long for the time allowed.

____19. This instructor can be expected to have a lot of questions pertaining to subject matter that was barely even mentioned in class or not mentioned at all.

____20. This instructor can be expected to give brief written exams which are corrected immediately.

____21. This instructor can be expected to have tests that require rote memorization.

____22. This instructor can be expected to avoid unclear test questions most of the time.

Appendix III

Behavioral Expectation Scales (BES)

FORM A

The following pages include nine categories concerning various aspects of teaching with definitions provided for each one. The categories and definitions resulted directly from meetings held with groups of college students. Below each category is a 7-point rating scale with behavioral statements located at various points on the scale. All of the statements were written by college students and each statement's location on a scale was determined by student evaluations of the level of performance it best represented.

Please read the definition for the first category and the behavioral statements shown on the right of the accompanying scale. Then compare those behaviors of your instructor which apply to this category against the behaviors on the scale. Finally, use the behaviors on the scale as references or aides in helping you determine the rating that you feel your instructor deserves. That is, based upon the behaviors you have seen, rate your instructor according to the type of behavior that you would expect of him/her when compared to the behaviors on the scale. It is possible for your instructor to have never exhibited any of the behaviors shown on the scale. However, based upon those behaviors that you have seen exhibited by your instructor which are applicable to the category being rated, rate your instructor on the level of performance you would expect from him/her relative to the behavioral statements on the scale.

Please make a single mark anywhere on the vertically positioned line. It is not necessary to mark at a point where a statement is located. The statements are only references against which you compare your observed behaviors of the instructor.

At the bottom of the page is a 5-point scale ranging from very uncertain(1) to very certain(5). After rating your instructor on this category, then decide how sure you are of this rating and circle the one number(either the 1, 2, 3, 4, or 5) which best describes how certain you are of your rating.

Rate your instructor on the remaining categories using the same strategy.

ASSIGNMENTS: Extent to which the instructor is clear on what is to be done, avoids assigning excessive amounts, and provides assignments which contribute to the understanding of the subject matter rather than just providing busy work.

7

Can be expected to clearly explain what is to be done on an assignment and assign enough work to understand the concepts to be tested and needed later.

6

Can be expected to give plenty of time for students to get assignments done.

5

Can be expected to assign questions that do not have to be turned in, but still discusses them in class.

4

Can be expected to assign too much work once in a while.

3

Can be expected to assign very difficult homework.

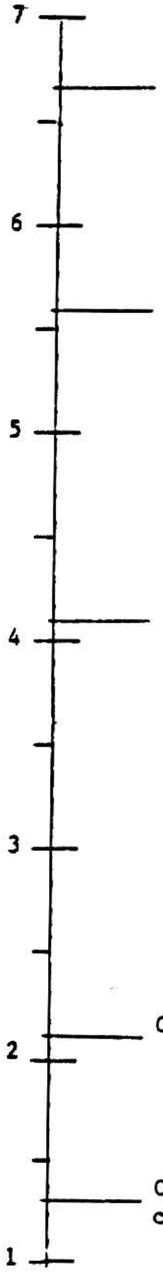
2

Can be expected to give only busy work.

1

Certainty of Rating: 1 2 3 4 5
Very Uncertain — Very Certain

ATTITUDE TOWARDS SUBJECT: Extent to which the instructor shows personal interest in the material and displays a positive attitude towards teaching the subject.



Can be expected to be really excited about what is taught.

Can be expected to tell students that he/she enjoys the work.

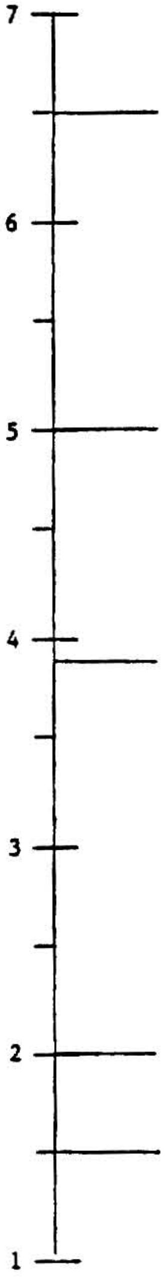
Can be expected to come each day and seem to enjoy presenting the material, yet look anxious to leave upon dismissal of class.

Can be expected to be late for class many times.

Can be expected to show no interest and act like he/she does not care for the subject at all.

Certainty of Rating: 1 2 3 4 5
Very Uncertain ... Very Certain

<u>GRADES</u>: Extent to which the instructor's grading practices remain consistent and free of confusion and are also fair.



Can be expected to explain his/her grading method so the students know how grades are determined and also have evidence for the grades.

Can be expected to give grades based on test scores and homework.

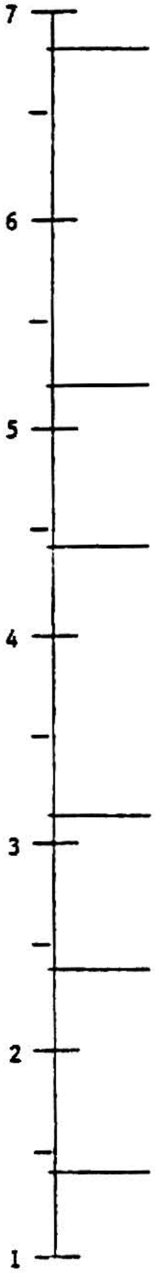
Can be expected to grade on a percentage of total possible with A = 90-100; B = 80-89; C = 70-79; D = 60-69; and F = below 60.

Can be expected to grade with such diversity that questions are always asked at the time of grades.

Can be expected to grade in a way that favors a particular kind of student.

Certainty of Rating:	1	2	3	4	5
	Very Uncertain				Very Certain

INSTRUCTOR KNOWLEDGE: Extent to which the instructor is aware of current material related to the course or to his/her field and is able to accurately answer or direct the student to specific sources that will answer questions concerning the subject matter.



Can be expected to be up to date in the field and able to relate new concepts and ideas to the text and answer questions concerning them.

Can be expected to find answers to most of the students' questions and the answers are fairly accurate.

Can be expected to give a student some idea of where to go to find an answer to a question.

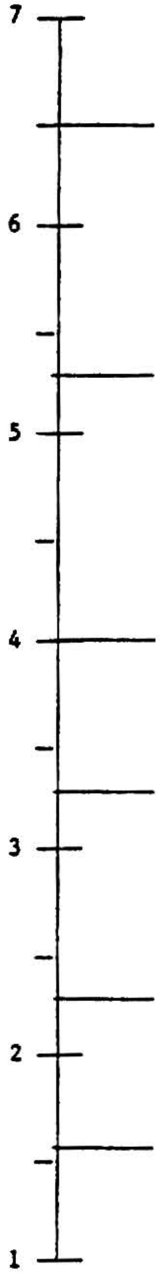
Can be expected to answer questions but rely on rather old material.

Can be expected to know very little of current research efforts in his/her field.

Can be expected to beat around the bush when a student asks a question and does not seem to really know the answer.

Certainty of Rating: 1 2 3 4 5
Very Uncertain — Very Certain

<u>MANNER OF PRESENTATION</u>: Extent to which the instructor's methods of presentation and use of audio-visual aids help emphasize and clarify important points; ability to present material clearly and concisely on a level students can understand.



Can be expected to have actual demonstrations in class so the students can experience what they are studying and also have films to illustrate points.

Can be expected to have guest speakers every once in a while.

Can be expected to lecture mostly and show a film every once in a while.

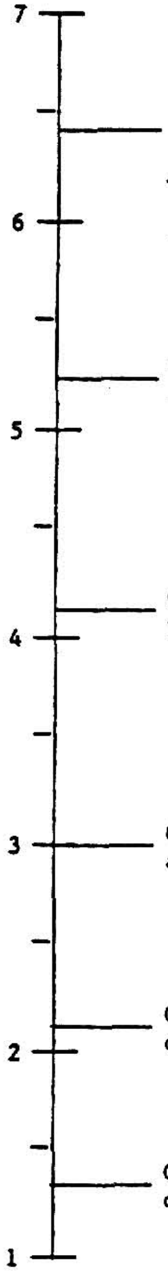
Can be expected to show films; however, they are usually very old.

Can be expected to rarely, if ever, use audio-visual aid.

Can be expected to not convey the material on a level students can understand.

Certainty of Rating: 1 2 3 4 5
Very Uncertain — Very Certain

OBJECTIVENESS: Extent to which the instructor remains objective and presents a fair treatment of all points of view on controversial or debatable topics.



Can be expected to openly discuss each side of a debatable topic and not let personal feelings interfere, yet still offer his/her knowledge readily.

Can be expected to listen to the students' points of view and also make his/her own point of view known.

Can be expected to clearly show favor for a certain side, but still present the other side.

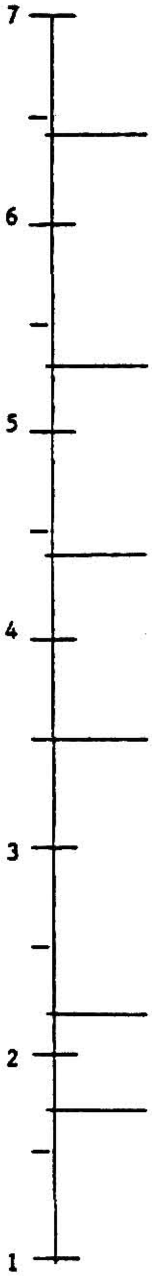
Can be expected to listen to the student's point of view and then argue against it.

Can be expected to be biased and, although saying he/she welcomes debate, becomes very defensive when questions are raised.

Can be expected to act as if his/her views are the only correct ones and not listen to other views.

Certainty of Rating: 1 2 3 4 5
Very Uncertain — Very Certain

<u>ORGANIZATION</u>: Extent to which the instructor arranges the subject matter and course objectives in an orderly and logical sequence for thorough coverage.



Can be expected to hand out a schedule for the semester which shows objectives for tests as well as outside material to be covered in lectures.

Can be expected to verbally present and explain a general course outline on the first day of class.

Can be expected to get a little confused but on the whole has what he/she is going to say fairly well planned.

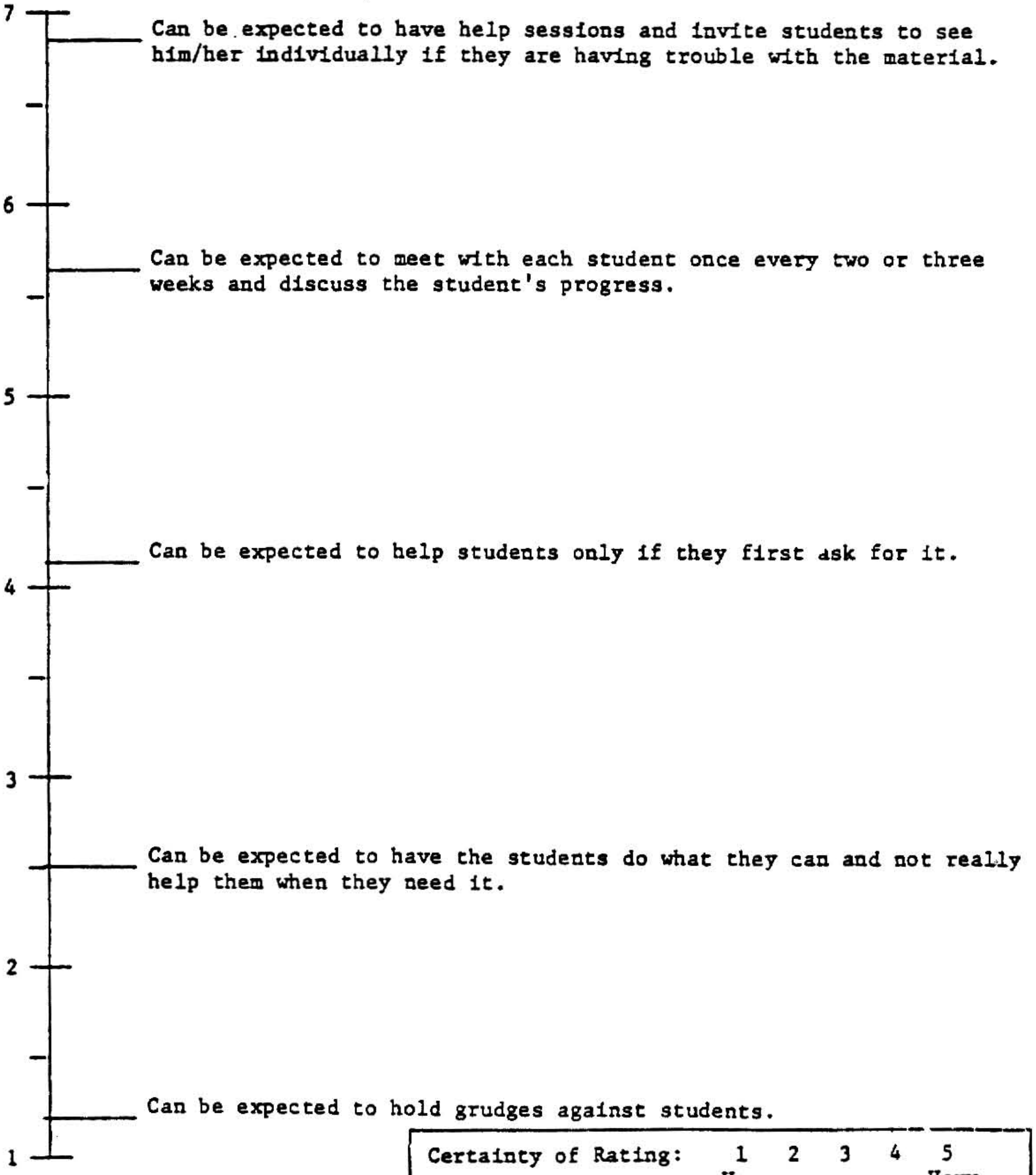
Can be expected to have some organization but, at times, get off the topic and discuss something totally unrelated.

Can be expected to start talking about the subject without well planned direction and ends up repeating things and getting a little lost.

Can be expected to jump all around in the subject and students find it very difficult to follow.

Certainty of Rating: 1 2 3 4 5
Very Uncertain — Very Certain

STUDENT-TEACHER RELATIONS: Extent to which the instructor shows a true, sincere concern for the welfare of the students through such things as dependability, availability for help, and consideration of student feelings; establishing rapport with the students.



Can be expected to have help sessions and invite students to see him/her individually if they are having trouble with the material.

Can be expected to meet with each student once every two or three weeks and discuss the student's progress.

Can be expected to help students only if they first ask for it.

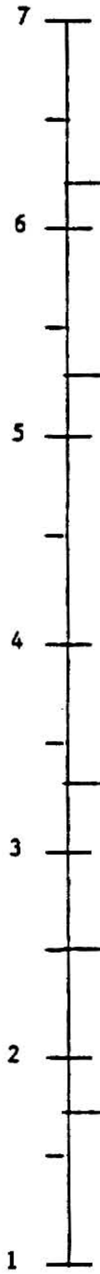
Can be expected to have the students do what they can and not really help them when they need it.

Can be expected to hold grudges against students.

Certainty of Rating: 1 2 3 4 5

Very Uncertain — Very Certain

TESTS: Extent to which the instructor writes clear, unambiguous questions that relate to and are representative of in-class material and outside readings which were stressed adequately in class.



Can be expected to test the class on relevant material that has been discussed in class.

Can be expected to give brief written exams which are corrected immediately.

Can be expected to test more heavily on some parts of the material than on others without telling the students.

Can be expected to occasionally test on material not covered in class.

Can be expected to give tests that are too long for the time allowed.

Certainty of Rating:	1 Very Uncertain	2	3	4	5 Very Certain

FORM B

The following pages include nine categories concerning various aspects of teaching with definitions provided for each one. The categories and definitions resulted directly from meetings held with groups of college students. Below each category is a 7-point rating scale with behavioral statements located at various points on the scale. All of the statements were written by college students and each statement's location on a scale was determined by student evaluations of the level of performance it best represented.

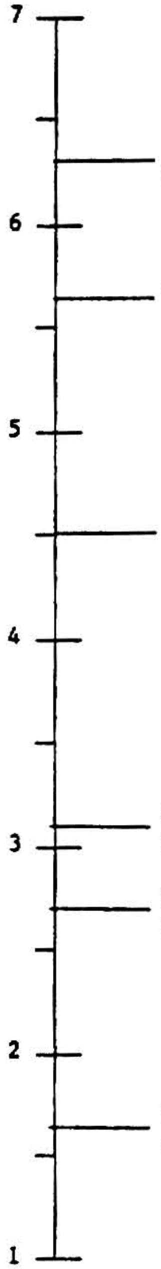
Please read the definition for the first category and the behavioral statements shown on the right of the accompanying scale. Then compare those behaviors of your instructor which apply to this category against the behaviors on the scale. Finally, use the behaviors on the scale as references or aides in helping you determine the rating that you feel your instructor deserves. That is, based upon the behaviors you have seen, rate your instructor according to the type of behavior that you would expect of him/her when compared to the behaviors on the scale. It is possible for your instructor to have never exhibited any of the behaviors shown on the scale. However, based upon those behaviors that you have seen exhibited by your instructor which are applicable to the category being rated, rate your instructor on the level of performance you would expect from him/her relative to the behavioral statements on the scale.

Please make a single mark anywhere on the vertically positioned line. It is not necessary to mark at a point where a statement is located. The statements are only references against which you compare your observed behaviors of the instructor.

At the bottom of the page is a 5-point scale ranging from very uncertain(1) to very certain(5). After rating your instructor on this category, then decide how sure you are of this rating and circle the one number(either the 1, 2, 3, 4, or 5) which best describes how certain you are of your rating.

Rate your instructor on the remaining categories using the same strategy.

ASSIGNMENTS: Extent to which the instructor is clear on what is to be done, avoids assigning excessive amounts, and provides assignments which contribute to the understanding of the subject matter rather than just providing busy work.



Can be expected to give assignments that relate to the subject matter and provide further understanding of it.

Can be expected to tell students which problems in the assignment are more important.

Can be expected to give assignments that offer at least some understanding of the subject matter.

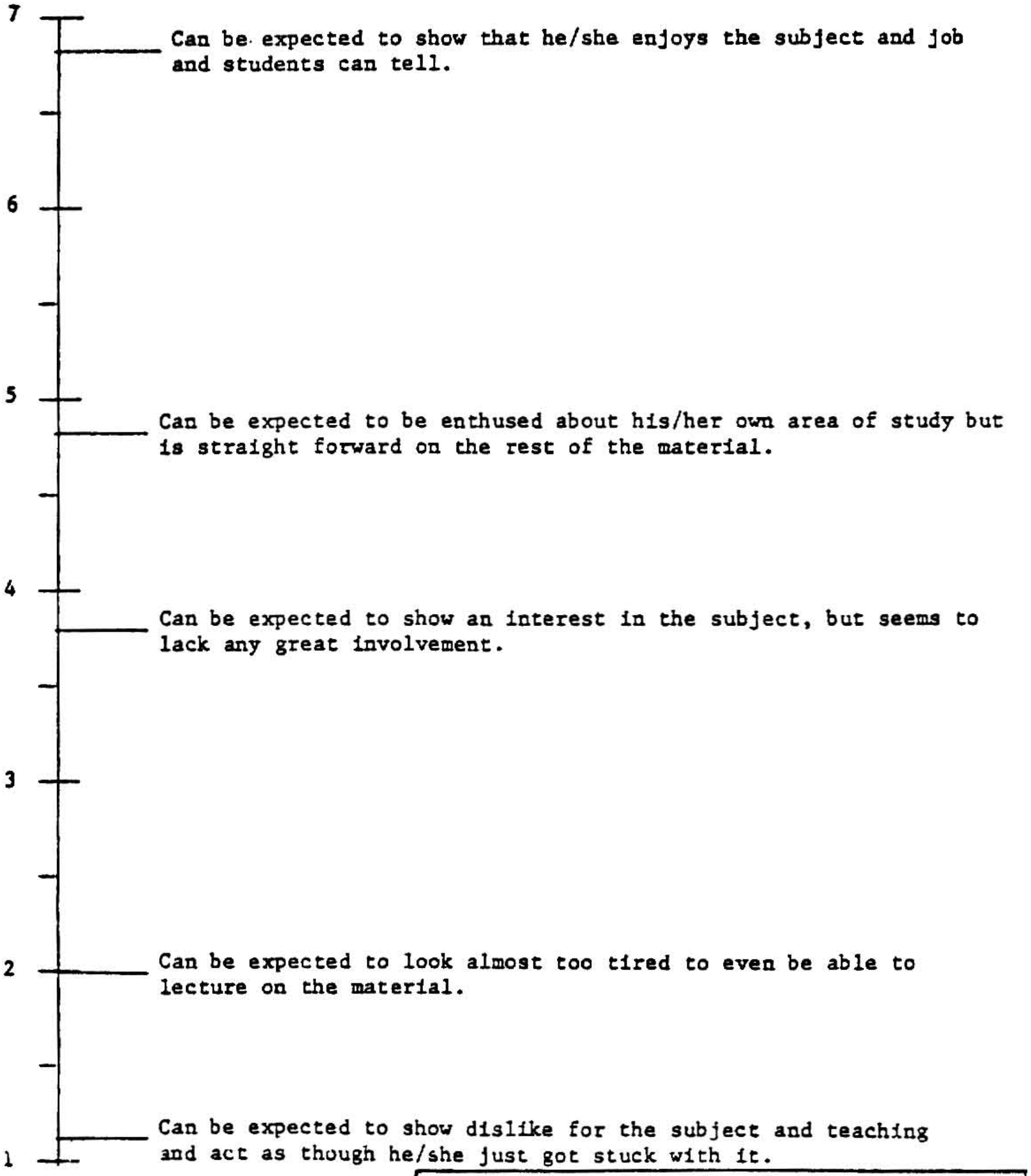
Can be expected to give assignments that cover material much deeper than the students need to learn.

Can be expected to assign a lot of problems that seem to cover the same thing.

Can be expected to vaguely give the student an idea of what is to be done on the next assignment and not really say when it's due.

Certainty of Rating: 1 2 3 4 5
Very Uncertain ... Very Certain

ATTITUDE TOWARDS SUBJECT: Extent to which the instructor shows personal interest in the material and displays a positive attitude towards teaching the subject.



Can be expected to show that he/she enjoys the subject and job and students can tell.

Can be expected to be enthused about his/her own area of study but is straight forward on the rest of the material.

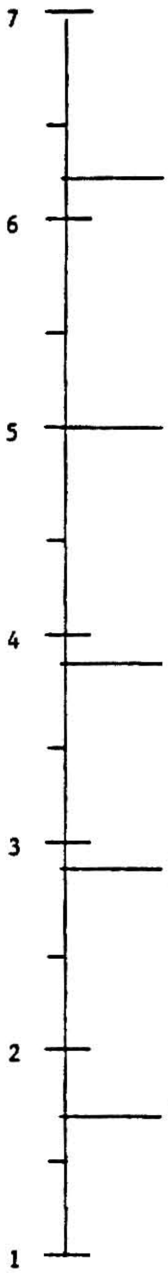
Can be expected to show an interest in the subject, but seems to lack any great involvement.

Can be expected to look almost too tired to even be able to lecture on the material.

Can be expected to show dislike for the subject and teaching and act as though he/she just got stuck with it.

Certainty of Rating:	1 Very Uncertain	2	3	4	5 Very Certain

<u>GRADES</u>: Extent to which the instructor's grading practices remain consistent and free of confusion and are also fair.



Can be expected to curve the grades according to the difficulty of the test and the overall student scores.

Can be expected to grade with a "C" as average.

Can be expected to use only test scores for determining grades.

Can be expected to maintain high standards even when the grades of the entire class are low in relation to that standard.

Can be expected to not let students know how grades are derived.

Certainty of Rating:	1 Very Uncertain	2	3	4	5 Very Certain

INSTRUCTOR KNOWLEDGE: Extent to which the instructor is aware of current material related to the course or to his/her field and is able to accurately answer or direct the student to specific sources that will answer questions concerning the subject matter.

7

Can be expected to add current material to class presentations for use as new examples.

6

Can be expected to be aware of new concepts and although unable to answer specific questions, can direct the student to sources that can.

5

Can be expected to inform students of fairly recent developments in the area during the past few years, but is not up to date on present studies.

4

3

Can be expected to have to continually refer back to notes to answer questions.

2

Can be expected to seem not to know what he/she is talking about and can't answer questions.

1

Certainty of Rating:	1 Very Uncertain	2	3	4	5 Very Certain

MANNER OF PRESENTATION: Extent to which the instructor's methods of presentation and use of audio-visual aids help emphasize and clarify important points; ability to present material clearly and concisely on a level students can understand.

7

Can be expected to use interesting films and discuss them afterwards in order to clarify major points.

6

Can be expected to do experiments to get a point across.

5

4

Can be expected to leave the lecture material only occasionally to give an interesting sideview.

3

Can be expected to use audio-visual aids but not clearly explain how it pertains to the lecture.

2

Can be expected to go through the material very fast and allow no time for questions.

Certainty of Rating:	1	2	3	4	5
	Very Uncertain				Very Certain

1

ILLEGIBLE DOCUMENT

THE FOLLOWING DOCUMENT(S) IS OF POOR LEGIBILITY IN THE ORIGINAL

THIS IS THE BEST COPY AVAILABLE

OBJECTIVENESS: Extent to which the instructor remains objective and presents a fair treatment of all points of view on controversial or debatable topics.

7

Can be expected to show the various points of view, allow discussion, not put down anyone, listen, and let students decide for themselves.

6

Can be expected to present both sides of a debatable topic in equal detail and without prejudice even if leaning more towards one particular view.

5

Can be expected to always state his/her point of view, although he/she will allow discussion of other points of view.

4

Can be expected to discuss each side of a controversial topic, but leans towards his/her beliefs and tries to convince others of them.

3

Can be expected to present both sides of the issue and then ignore the side he/she does not favor.

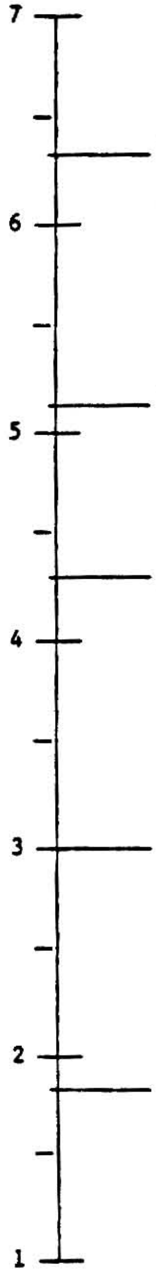
2

Can be expected to be completely close-minded and students that openly disagree with him/her have a difficult time in the class.

1

Certainty of Rating:	1	2	3	4	5
	Very Uncertain				Very Certain

ORGANIZATION: Extent to which the instructor arranges the subject matter and course objectives in an orderly and logical sequence for thorough coverage.



Can be expected to have an outlined, detailed schedule of each day's activities to present to the students at the beginning, but it can be changed if necessary.

Can be expected to organize the material into increasing levels of difficulty.

Can be expected to present material that sometimes overlaps with material that has already been presented.

Can be expected to jump back and forth from one chapter to another in the book.

Can be expected to give a vague sheet of chapter order and then rarely cover the material on schedule or give tests on projected days.

Certainty of Rating: 1 2 3 4 5
Very Uncertain — Very Certain

STUDENT-TEACHER RELATIONS: Extent to which the instructor shows a true, sincere concern for the welfare of the students through such things as dependability, availability for help, and consideration of student feelings; establishing rapport with the students.

7

Can be expected to invite students to come and talk with him/her after class or during office hours whenever they have a problem with the class.

6

Can be expected to recognize students who are in his/her classes.

5

4

3 Can be expected to have a student come to his/her office after class or wait until the next class for an answer to a question.

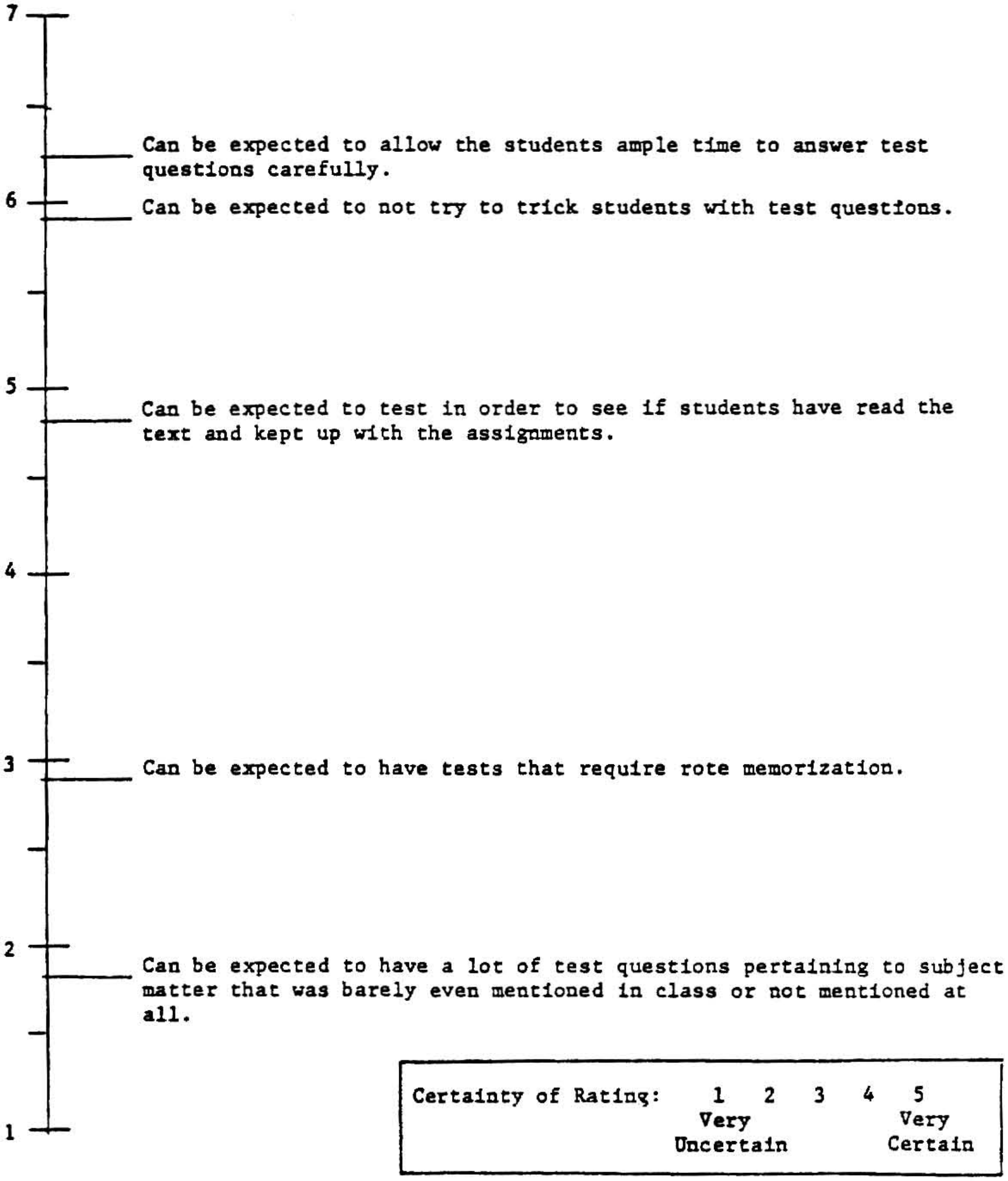
2 Can be expected to never have help sessions.

Can be expected to not even care if students do not understand the material.

1

Certainty of Rating:	1 Very Uncertain	2	3	4	5 Very Certain

TESTS: Extent to which the instructor writes clear, unambiguous questions that relate to and are representative of in-class material and outside readings which were stressed adequately in class.



Can be expected to allow the students ample time to answer test questions carefully.

Can be expected to not try to trick students with test questions.

Can be expected to test in order to see if students have read the text and kept up with the assignments.

Can be expected to have tests that require rote memorization.

Can be expected to have a lot of test questions pertaining to subject matter that was barely even mentioned in class or not mentioned at all.

Certainty of Rating:	1 Very Uncertain	2	3	4	5 Very Certain

Appendix IV

Mixed Standard Scales (MSS)

FORM A

Listed on the following pages are a number of descriptions of behavior concerning various aspects of teaching. To the right of each behavioral statement are three boxes. The box to the left corresponds with "The instructor is worse than this statement."; the middle box corresponds with "This statement fits the instructor."; and the box to the right corresponds with "The instructor is better than this statement." Carefully examine each statement and determine in your own mind the answer to the following question: Is the instructor I am rating "better" than this statement, "worse" than this statement, or does this statement "fit" the instructor?

If you believe the instructor's behavior is "better" than the behavior described in the statement, mark an "X" in the box to the right of the statement that corresponds with "The instructor is better than this statement." If you believe the instructor's behavior is "worse" than the behavior described in the statement, mark an "X" in the box that corresponds with "The instructor is worse than this statement." If you believe the statement "fits" the instructor you are rating, mark an "X" in the middle box to the right of the statement that corresponds with "This statement fits the instructor."

At the top of each page is a 5-point scale ranging from very uncertain (1) to very certain(5). After rating your instructor on a statement, then decide how sure you are of this rating and write the <u>one</u> number(either a 1, 2, 3, 4, or 5) which best describes how certain you are of your rating in the space to the left of the behavioral statement.

Certainty of Rating

1	2	3	4	5
Very Uncertain				Very Certain

Certainty Rating		Statements		Ratings		
				The instructor is worse than this statement.	This statement fits the instructor.	The instructor is better than this statement.
_____	1.	Can be expected to assign too much work once in a while.	1.			
_____	2.	Can be expected to jump all around in the subject and students find it very difficult to follow.	2.			
_____	3.	Can be expected to have actual demonstrations in class so the students can experience what they are studying and also have films to illustrate points.	3.			
_____	4.	Can be expected to act as if his/her views are the only correct ones and not listen to other views.	4.			
_____	5.	Can be expected to get a little confused, but on the whole has what he/she is going to say fairly well planned.	5.			
_____	6.	Can be expected to have help sessions and invite students to see him/her individually if they are having trouble with the material.	6.			
_____	7.	Can be expected to come each day and seem to enjoy presenting the material, yet look anxious to leave upon dismissal of class.	7.			
_____	8.	Can be expected to explain his/her grading method so the students know how grades are determined and also have evidence for the grades.	8.			
_____	9.	Can be expected to grade in a way that favors a particular kind of student.	9.			
_____	10.	Can be expected to beat around the bush when a student asks a question and does not seem to really know the answer.	10.			

Certainty of Rating

1 Very Uncertain 2 3 4 5 Very Certain

Certainty Rating		Statements		Ratings		
				The instructor is worse than this statement.	This statement fits the instructor.	The instructor is better than this statement.
_____	11.	Can be expected to clearly explain what is to be done on an assignment and assign enough work to understand the concepts to be tested and needed later.	11.			
_____	12.	Can be expected to test more heavily on some parts of the material than on others without telling the students.	12.			
_____	13.	Can be expected to be up to date in the field and able to relate new concepts and ideas to the text and answer questions concerning them.	13.			
_____	14.	Can be expected to not convey the material on a level students can understand.	14.			
_____	15.	Can be expected to lecture mostly and show a film every once in a while.	15.			
_____	16.	Can be expected to openly discuss each side of a debatable topic and not let personal feelings interfere, yet still offer his/her knowledge readily.	16.			
_____	17.	Can be expected to hold grudges against students.	17.			
_____	18.	Can be expected to be really excited about what is taught.	18.			
_____	19.	Can be expected to help students only if they first ask for it.	19.			
_____	20.	Can be expected to give only busy work.	20.			

Certainty of Rating

1 2 3 4 5

Very Uncertain — Very Certain

Certainty Rating		Statements		Ratings		
				The instructor is worse than this statement.	This statement fits the instructor.	The instructor is better than this statement.
_____	21.	Can be expected to clearly show favor for a certain side, but still present the other side.	21.			
_____	22.	Can be expected to grade on a percentage of total possible with A = 90-100; B = 80-89; C = 70-79; D = 60-69; and F = below 60.	22.			
_____	23.	Can be expected to show no interest and act like he/she does not care for the subject at all.	23.			
_____	24.	Can be expected to test the class on relevant material that has been discussed in class.	24.			
_____	25.	Can be expected to hand out a schedule for the semester which shows objectives for tests as well as outside material to be covered in lectures.	25.			
_____	26.	Can be expected to give a student some idea of where to go to find an answer to a question.	26.			
_____	27.	Can be expected to give tests that are too long for the time allowed.	27.			

FORM B

Listed on the following pages are a number of descriptions of behavior concerning various aspects of teaching. To the right of each behavioral statement are three boxes. The box to the left corresponds with "The instructor is worse than this statement."; the middle box corresponds with "This statement fits the instructor."; and the box to the right corresponds with "The instructor is better than this statement." Carefully examine each statement and determine in your own mind the answer to the following question: Is the instructor I am rating "better" than this statement, "worse" than this statement, or does this statement "fit" the instructor?

If you believe the instructor's behavior is "better" than the behavior described in the statement, mark an "X" in the box to the right of the statement that corresponds with "The instructor is better than this statement." If you believe the instructor's behavior is "worse" than the behavior described in the statement, mark an "X" in the box that corresponds with "The instructor is worse than this statement." If you believe the statement "fits" the instructor you are rating, mark an "X" in the middle box to the right of the statement that corresponds with "This statement fits the instructor."

At the top of each page is a 5-point scale ranging from very uncertain (1) to very certain(5). After rating your instructor on a statement, then decide how sure you are of this rating and write the one number(either a 1, 2, 3, 4, or 5) which best describes how certain you are of your rating in the space to the left of the behavioral statement.

Certainty of Rating

1	2	3	4	5
Very Uncertain				Very Certain

				Ratings		
Certainty Rating		Statements		The instructor is worse than this statement.	This statement fits the instructor.	The instructor is better than this statement.
_____	1.	Can be expected to give assignments that offer at least some understanding of the subject matter.	1.			
_____	2.	Can be expected to give a vague sheet of chapter order and then rarely cover the material on schedule or give tests on projected days.	2.			
_____	3.	Can be expected to use interesting films and discuss them afterwards in order to clarify major points.	3.			
_____	4.	Can be expected to be completely close-minded and students that openly disagree with him/her have a difficult time in the class.	4.			
_____	5.	Can be expected to present material that sometimes overlaps with material that has already been presented.	5.			
_____	6.	Can be expected to invite students to come and talk with him/her after class or during office hours whenever they have a problem with the class.	6.			
_____	7.	Can be expected to show an interest in the subject, but seem to lack any great involvement.	7.			
_____	8.	Can be expected to curve the grades according to the difficulty of the test and the overall student scores.	8.			
_____	9.	Can be expected to not let students know how grades are derived.	9.			
_____	10.	Can be expected to seem not to know what he/she is talking about and can't answer questions.	10.			

Certainty of Rating

1	2	3	4	5
Very Uncertain				Very Certain

Certainty Rating **Statements**

_____ 11. Can be expected to give assignments that relate to the subject matter and provide further understanding of it.

_____ 12. Can be expected to test in order to see if students have read the text and kept up with the assignments.

_____ 13. Can be expected to add current material to class presentations for use as new examples.

_____ 14. Can be expected to go through the material very fast and allow no time for questions.

_____ 15. Can be expected to leave the lecture material only occasionally to give an interesting sideview.

_____ 16. Can be expected to show the various points of view, allow discussion, not put down anyone, listen, and let students decide for themselves.

_____ 17. Can be expected to not even care if students do not understand the material.

_____ 18. Can be expected to show that he/she enjoys the subject and job and students can tell.

_____ 19. Can be expected to have a student come to his/her office after class or wait until the next class for an answer to a question.

_____ 20. Can be expected to vaguely give the student an idea of what is to be done on the next assignment and not really say when it's due.

Ratings

	The instructor is worse than this statement.	This statement fits the instructor.	The instructor is better than this statement.
11.			
12.			
13.			
14.			
15.			
16.			
17.			
18.			
19.			
20.			

Certainty of Rating

1 2 3 4 5

Very Uncertain — Very Certain

Certainty Rating — Statements

_____ 21. Can be expected to discuss each side of a controversial topic but lean towards his/her beliefs and try to convince others of them.

_____ 22. Can be expected to use only test scores for determining grades.

_____ 23. Can be expected to show dislike for the subject and teaching and act as though he/she just got stuck with it.

_____ 24. Can be expected to allow the students ample time to answer test questions carefully.

_____ 25. Can be expected to have an outlined, detailed schedule of each day's activities to present to the students at the beginning, but it can be changed if necessary.

_____ 26. Can be expected to inform students of fairly recent developments in the area during the past few years, but is not up to date on present studies.

_____ 27. Can be expected to have a lot of test questions pertaining to subject matter that was barely even mentioned in class or not mentioned at all.

Ratings

	The instructor is worse than this statement.	This statement fits the instructor.	The instructor is better than this statement.
21.			
22.			
23.			
24.			
25.			
26.			
27.			

Appendix V

Numerical Values of Response Combinations for Original and Revised Coding Systems for Mixed Standard Scales

Original and Revised Numerical Ratings for the 27 Possible Mixed Standard Scale Response Combinations

	Response Combinations:			Numerical Ratings:	
	Superior Behavior	Average Behavior	Inferior Behavior	Original	Revised
1.	+	+	+	7	7
2.	+	+	0	7	6
3.	+	+	-	7	5
4.	+	0	+	4	6
5.	+	0	0	3	5
6.	+	0	-	4	4
7.	+	-	+	3	5
8.	+	-	0	2	4
9.	+	-	-	1	3
10.	0	+	+	6	6
11.	0	+	0	6	5
12.	0	+	-	6	4
13.	0	0	+	omitted	5
14.	0	0	0	4	4
15.	0	0	-	4	3
16.	0	-	+	5	4
17.	0	-	0	2	3
18.	0	-	-	1	2

	Response Combinations:			Numerical Ratings:	
	Superior Behavior	Average Behavior	Inferior Behavior	Original	Revised
19.	-	+	+	5	5
20.	-	+	0	5	4
21.	-	+	-	5	3
22.	-	0	+	4	4
23.	-	0	0	omitted	3
24.	-	0	-	3	2
25.	-	-	+	3	3
26.	-	-	0	2	2
27.	-	-	-	1	1

Note. + indicates "ratee is better than this behavior."

0 indicates "ratee is the same as this behavior."

- indicates "ratee is worse than this behavior."

REVISED CODING SYSTEM

	Performance Levels of Behavioral Descriptions:		
	Superior	Average	Inferior
Possible	+	+	+
MSS Responses:	0	0	0
	-	-	-
Numerical	8	5	2
Equivalents:	7	4	1
	6	3	0

Numerical Rating Assigned to Response Combination =

Numerical Equivalent of "Superior" Response (8, 7, or 6)

+ Numerical Equivalent of "Average" Response (5, 4, or 3)

+ Numerical Equivalent of "Inferior" Response (2, 1, or 0)

- 8 (a constant)

Note. + indicates "ratee is better than this behavior."

0 indicates "ratee is the same as this behavior."

- indicates "ratee is worse than this behavior."

Appendix VI

Graphic Scales (GRS)

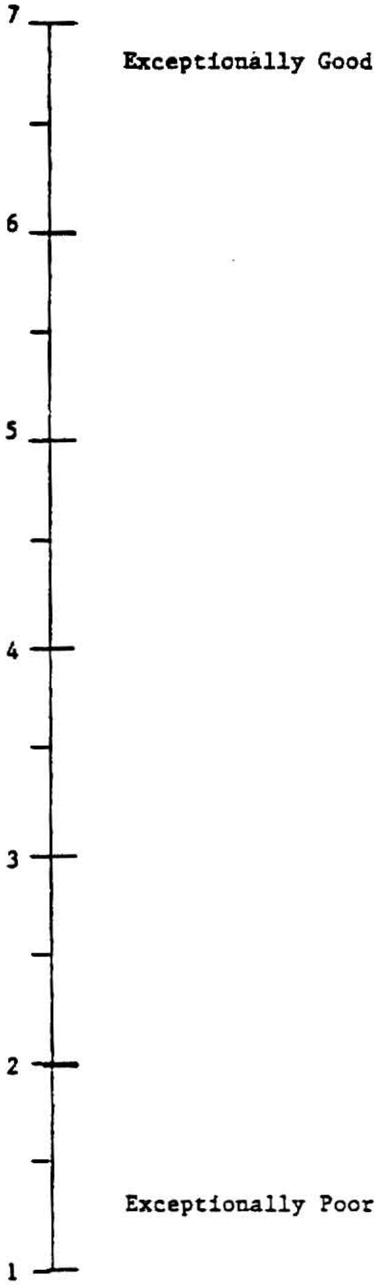
The following pages include nine categories concerning various aspects of teaching with definitions provided for each one. The categories and definitions resulted directly from meetings held with groups of college students. Below each category is a 7-point rating scale.

Please read the definition for the first category. Considering only those behaviors of your instructor that apply to this category, now determine the rating that you feel your instructor deserves. Then make a single mark on the vertically positioned line which reflects your assessment of his/her performance.

At the bottom of the page is a 5-point scale ranging from very uncertain(1) to very certain(5). After rating your instructor on this category, then decide how sure you are of this rating and circle the one number(either the 1, 2, 3, 4, or 5) which best describes how certain you are of this rating.

Rate your instructor on the remaining categories using the same strategy.

<u>ASSIGNMENTS</u>: Extent to which the instructor is clear on what is to be done, avoids assigning excessive amounts, and provides assignments which contribute to the understanding of the subject matter rather than just providing busy work.

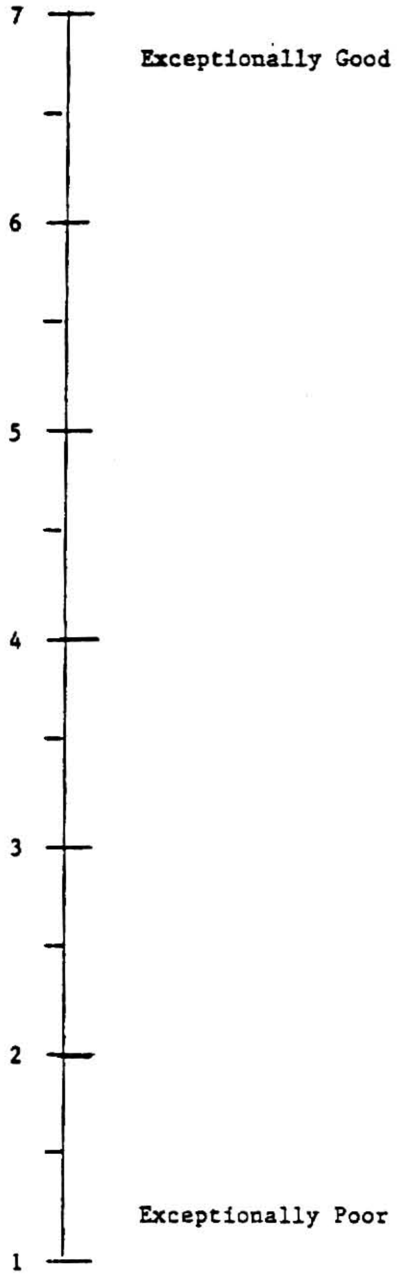


<u>Certainty</u> <u>of</u> <u>Rating</u>

1 2 3 4 5

Very Uncertain — Very Certain

ATTITUDE TOWARDS SUBJECT: Extent to which the instructor shows personal interest in the material and displays a positive attitude towards teaching the subject.

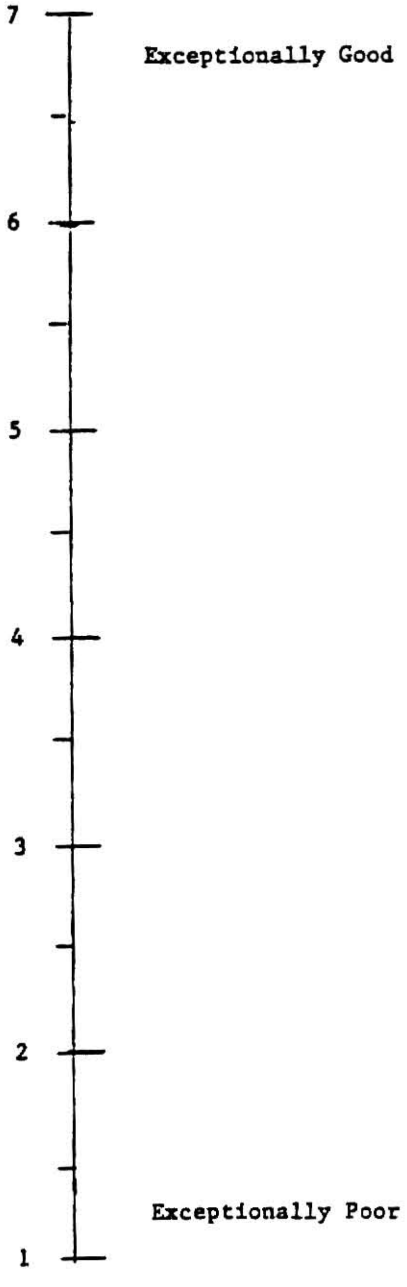


Certainty of Rating

1 2 3 4 5

Very Uncertain — Very Certain

GRADES: Extent to which the instructor's grading practices remain consistent and free of confusion and are also fair.



Certainty	of	Rating		
1	2	3	4	5
Very Uncertain				Very Certain

INSTRUCTOR KNOWLEDGE: Extent to which the instructor is aware of current material related to the course or to his/her field and is able to accurately answer or direct the student to specific sources that will answer questions concerning the subject matter.

7

Exceptionally Good

6

5

4

3

2

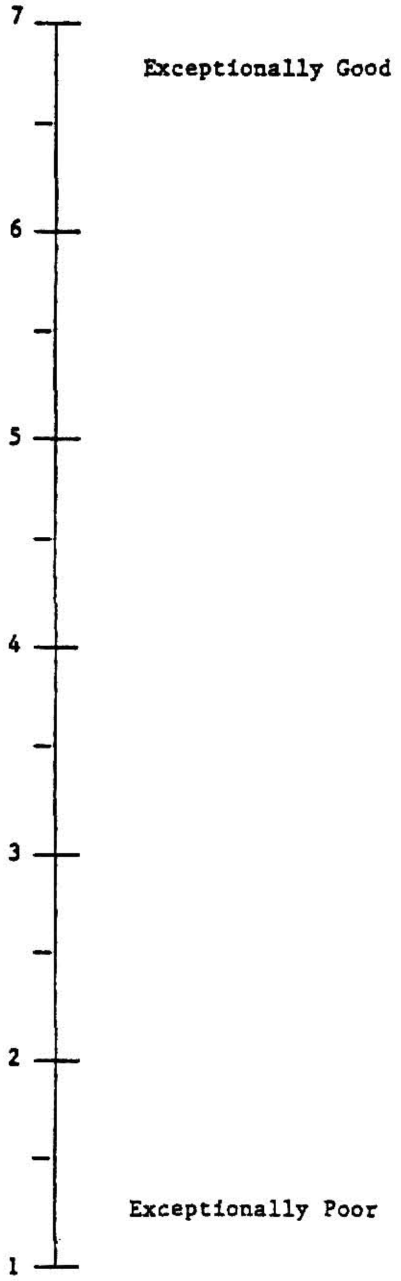
Exceptionally Poor

1

Certainty of Rating

1	2	3	4	5
Very Uncertain				Very Certain

MANNER OF PRESENTATION: Extent to which the instructor's methods of presentation and use of audio-visual aids help emphasize and clarify important points; ability to present material clearly and concisely on a level students can understand.



Certainty	of	Rating		
1	2	3	4	5
Very Uncertain				Very Certain

OBJECTIVENESS: Extent to which the instructor remains objective and presents a fair treatment of all points of view on controversial or debatable topics.

7

Exceptionally Good

6

5

4

3

2

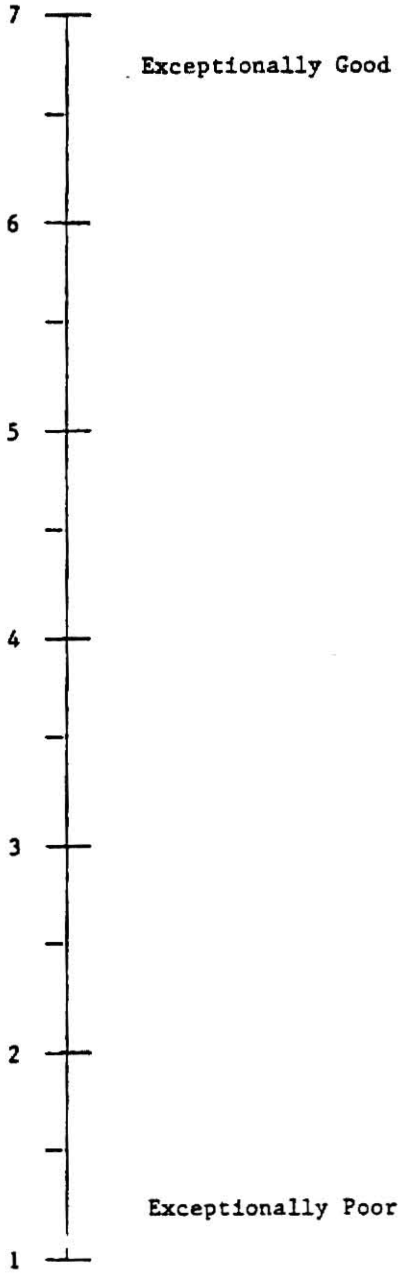
Exceptionally Poor

1

Certainty of Rating

1	2	3	4	5
Very Uncertain				Very Certain

ORGANIZATION: Extent to which the instructor arranges the subject matter and course objectives in an orderly and logical sequence for thorough coverage.



Certainty of Rating

1 2 3 4 5

Very Uncertain — Very Certàin

STUDENT-TEACHER RELATIONS: Extent to which the instructor shows a true, sincere concern for the welfare of the students through such things as dependability, availability for help, and consideration of student feelings; establishing rapport with the students.

7

Exceptionally Good

6

5

4

3

2

Exceptionally Poor

1

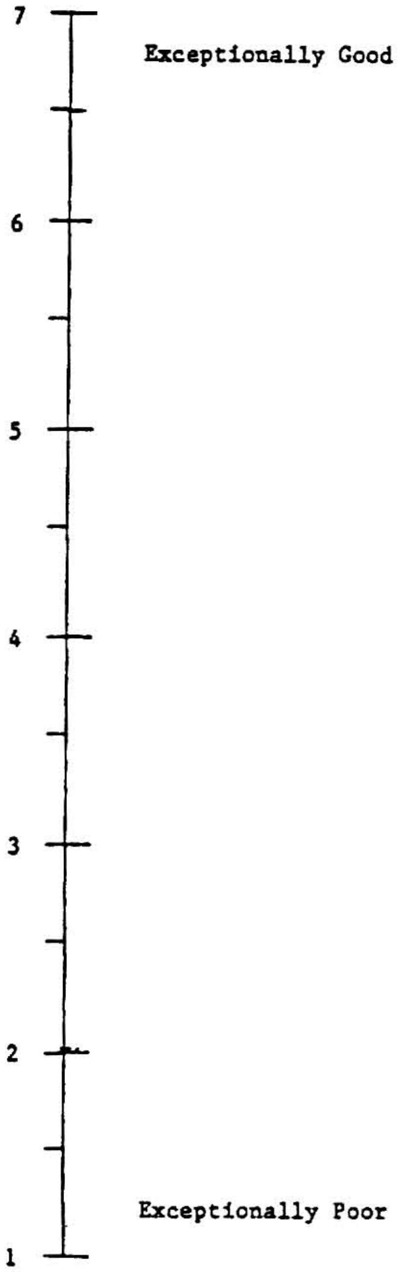
Certainty of Rating

1 2 3 4 5

Very Uncertain

Very Certain

__TESTS__: Extent to which the instructor writes clear, unambiguous questions that relate to and are representative of in-class material and outside readings which were stressed adequately in class.



__Certainty__ __of__ __Rating__

1 2 3 4 5

Very Very

Uncertain Certain

Appendix VII

Training Pamphlet

Performance Ratings

By now you have probably developed a rather general impression of your instructor which may be quite favorable or unfavorable. This impression may have been the direct result of a specific situation or it may have developed over a more prolonged period of time; it may have resulted from things that happened in class or from things that happened out of class, or even from a combination of the above. Yet, regardless of the cause, this general overall impression of the instructor has a direct effect on anything you do that pertains to him/her. In fact, in the area of performance rating, it is such a well known phenomenon that it has been given a special name. It's referred to as the "halo effect." For example, from several personal encounters with this instructor you discover that he/she is a very friendly, "easy going" kind of person whom you enjoy talking with and, as a result, you develop a generally favorable impression of him/her. Then when asked to evaluate his/her performance, you allow this general impression to influence your ratings and, rather than provide the instructor with accurate feedback regarding his/her teaching performance in each category, you proceed to give fairly high ratings on all of the categories which are consistent with your general impression of this instructor. This is commonly referred to as positive halo. Similarly, this general impression can result directly from an instructor's performance on some specific aspect of teaching which a student feels is extremely important. As another example, you may firmly believe it is most important for an instructor to give exams that test for understanding and not for ability to detect "delicately" worded phrases designed to trick the unsuspecting student. However, from the instructor's tests you find that the instructor seems to believe exactly the opposite. As a result of the instructor's testing practices you foster an unfavorable general impression of the instructor and thus consider him/her to be a poor teacher. If you allow this general impression, based upon his/her testing practices, to influence the ratings on every category, then the instructor will receive fairly low ratings on every category, consistent with your impression of the instructor as a poor teacher due to his/her testing practices. This is referred to as negative halo. One other possibility is that you may consider the instructor to be average and rate him/her as average on every category as opposed to another. Although it is entirely possible for an instructor to be very good in every aspect of teaching, average in every aspect of teaching, or very poor in every aspect of teaching, you should attempt to rate the instructor believing that this is not necessarily true.

The point of all this is not to change your overall impression of the instructor, but simply to inform you of the adverse effect of such an impression on ratings. What I would like you to do is try to distinguish among the various teaching categories and rate the instructor according to behavior that pertains only to that aspect of teaching and try not to let your ratings be influenced by some other unrelated behaviors or by this general overall impression of the instructor.

Another tendency of some raters is to avoid assigning extremely high or extremely low ratings to anybody and, instead, assign ratings that are closer to the middle of the scale.

On the other hand, some raters tend to give low negative ratings whereas other raters tend to give high positive ratings to anybody they rate. The first is "negative leniency" and the second is "positive leniency." The reason for informing you of these rater tendencies is simply to let you be aware of them and to encourage you to consider the full range of options on the scale when rating this instructor.

List of References

Reference Notes

1. Downey, R.G., & Saal, F.E. *Evaluating human judgment techniques*. Paper presented at the annual meeting of the American Psychological Association, Toronto, August, 1978.

2. Saal, F.E., Downey, R.G., and Lahey, M.A. *Rating the ratings: A multivariate approach to assessing the psychometric quality of rating data*. Unpublished manuscript, Kansas State University, 1979.

3. Blanz, F. *Mixed standard scale: A new merit rating method*. Unpublished doctoral dissertation, Helsinki, 1965.

4. Saal, F.E. *Development of job related rating scales for evaluating the performance of police patrol officers*. (Final Rep. KS78-01C, #11). Topeka, Kansas.

5. Kemp, K.E., & Dayton, A.D. *Alternative methods of testing equality of means compared by computer simulation*. Unpublished manuscript, Kansas State University, 1979.

References

Arvey, R.D., & Hoyle, J.C. A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer analysts. Journal of Applied Psychology, 1974, 59, 61-68.

Bernardin, H.J. Behavioral expectation scales versus summated scales: A fairer comparison. Journal of Applied Psychology, 1977, 62, 422-427.

Bernardin, H.J. Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 1978, 63, 301-308.

Bernardin, H.J., Alvares, K.M., & Cranny, C.J. A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 1976, 61, 564-570.

Bernardin, H.J., LaShells, M.B., Smith, P.C., & Alvares, K.M. Behavioral expectation scales: Effects of developmental procedures and formats. Journal of Applied Psychology, 1976, 61, 75-79.

Bernardin, H.J., & Walter, C.S. Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Blanz, F., & Ghiselli, E.E. The mixed standard scale: A new rating system. Personnel Psychology, 1972, 25, 185-199.

Blood, M.R. Spin-offs from behavioral expectation scale procedures. Journal of Applied Psychology, 1974, 59, 513-515.

Borman, W.C. The ratings of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 1974, 12, 105-124.

Borman, W.C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 1975, 60, 556-560.

Borman, W.C. Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 1979, 64. 410-421.

Borman, W.D., & Dunnette, M.D. Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 1975, 60, 561-565.

Borman, W.D., & Vallon, W.R. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. Journal of Applied Psychology, 1974, 59, 197-201.

Burnaska, R.F. & Hollmann, T.D. An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 59, 307-312.

Campbell, J. P., Dunnette, M.D., Arvey, R.D., & Hellervick, L.V. The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 1973, 57, 15-22.

Campbell, J.P., Dunnette, M.D., Lawler, E.E., & Weick, K.E. Managerial behavior, performance, and effectiveness. New York: McGraw-Hill, 1970.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validity by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Campion, J.E., Greener, J., & Vernli, S. Work observation versus recall in developing behavioral examples for rating scales. Journal of Applied Psychology, 1973, 58, 286-288.

Cascio, W.F., & Valenzi, E.R. Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. Journal of Applied Psychology, 1977, 62, 278-282.

Cronbach, L.J. The two disciplines of scientific psychology. The American Psychologist, 1957, 12, 671-684.

DeCotiis, T.A. An analysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 1977, 19, 247-266.

Dickinson, T.L., & Tice, T.E. A multitrait-multimethod analysis of scales developed by retranslation. Organizational Behavior and Human Performance, 1973, 9, 421-438.

Ellson, D.G., & Ellson, E.C. Historical note on the rating scale. Psychological Bulletin, 1953, 50, 383-384.

Finley, D.M., Osburn, H.G., Dubin, J.A., & Jeanneret, P.R. Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. Personnel Psychology, 1977, 30, 659-669.

Flanagan, J.C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-356.

Fogli, L., Hulin, C.L., & Blood, M.R. Development of first-level behavioral job criteria. Journal of Applied Psychology, 1971, 55, 3-8.

Friedman, B.A., & Cornelius III, E.T. Effect of rater participation in scale construction on the psychometric characteristics of two rating scale formats. Journal of Applied Psychology, 1976, 61, 210-216.

Ghiselli, E.E. Theory of Psychological Measurement. New York: McGraw-Hill, 1964.

Goodale, J.G., & Burke, R.J. Behaviorally based rating scales need not be job sepcific. Journal of Applied Psychology, 1975, 60, 389-391.

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

Guion, R. Personnel Testing. New York: McGraw-Hill, 1965.

Harari, O. & Zececk, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology, 1973, 58, 261-265.

Kafry, D., Zedeck, S., & Jacobs, R. The scalability of behavioral expectation scales as a function of developmental criteria. *Journal of Applied Psychology*, 1976, *61*, 519-522.

Kavanagh, M.J., MacKinney, A.C., & Walins, L. Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin*, 1971, *75*, 34-49.

Keaveny, T.J., & McGann, A.F. A comparieon of behavioral expectation scales and graphic rating scales. *Journal of Applied Psychology*, 1975, *60*, 695-703.

Klimoski, R.J., & London, M. Role of the rater in performance appraisal. *Journal of Applied Psychology*, 1974, *59*, 445-451.

Landy, F.J., & Guion, R.M. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, 1970, *5*, 93-103.

Landy, F.J., Farr, J.L., Saal, F.E., & Freytag, W.R. Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology*, 1976, *61*, 750-758.

Landy, F.J., & Trumbo, D. *Psychology of work behavior*. Homewood, Illinois: Dorsey Press, 1976.

Latham, G.P., Wexley, K.N., & Pursell, E.D. Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology*, 1975, *60*, 550-555.

Lawler, E.E., III. The multitrait-multirater appraoch to measuring managerial job performance. *Journal of Applied Psychology*, 1967, *51*, 369-381.

Mass, J.B. Patterned scaled expectation interview: Reliability studies on a new technique. *Journal of Applied Psychology*, 1965, *49*, 431-433.

Motovidlo, S.J., & Borman, W.D. Behaviorally anchored scales for measuring morale in military units. *Journal of Applied Psychology*, 1977, *62*, 177-183.

Saal, F.E., & Landy, F.J. The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance*, 1977, *18*, 19-35.

Saal, F.E. The mixed standard rating scale: A consistent system for numerically coding inconsistent response combinations. *Journal of Applied Psychology*, 1979, *64*, 422-428.

Schneier, C.E. Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. *Journal of Applied Psychology*, 1977, *62*, 541-548.

Schwab, D.P., Heneman III, H.G., & DeCotiis, T.A. Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology*, 1975, *28*, 549-562.

Seigel, S. *Nonparametric statistics: For the behavioral sciences*. New York: McGraw-Hill, 1965.

Smith, P.C. & Kendall, L.M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Vance, R.J., Kuhnert, K.W., & Farr, J.L. Interview judgments: Using external criteria to compare behavioral and graphic scale ratings. Organizational Behavior and Human Performance, 1978, 22, 279-295.

Warmke, D.L., & Billings, R.S. Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. Journal of Applied Psychology, 1949, 64, 124-131.

Zedeck, S., & Baker, H.T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. Organizational Behavior and Human Performance, 1972, 7, 457-466.

Zedeck, S., Imparato, N., Krausz, M., & Oleno, T. Development of behaviorally anchored rating scales as a function of organizational level. Journal of Applied Psychology, 1974, 59, 249-252.

Zedeck, S., Jacobs, R., & Kafry, D. Behavioral expectations: Development of parallel forms and analysis of scale assumptions. Journal of Applied Psychology, 1976, 61, 112=115.

Zedeck, S., Kafry, D., & Jacobs, R. Format and scoring variations in behavioral expectation evaluations. Organizational Behavior and Human Performance, 1976, 17, 171-184.

RATING SCALE FORMAT AND THE EFFECTIVENESS OF TRAINING RATERS TO MINIMIZE RATING ERRORS

by

LOREN K. KIRKEIDE

B. A., Minot State College, 1974

AN ABSTRACT OF A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Psychology

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1980

While research has failed to unequivocally demonstrate inherent psychometric strengths and weaknesses in different rating scale formats, studies examining rater training have generally shown that common rating errors (e.g., halo, leniency) can be reduced when raters are trained to avoid them. It was hypothesized that the efficacy of training raters to improve their ratings may be partially determined by the scale format. It was further proposed that differences in the relative psychometric properties of different rating scale formats may change considerably when the raters are subjected to a brief training session. Such training could conceivably eliminate psychometric differences among different rating scale formats that have been found in studies using untrained (and sometimes rather inexperienced) raters.

Two related objectives of this study were to (1) examine the possible impact of scale format on the effectiveness of a brief training session for raters and (2) compare the relative psychometric properties of different formats when using untrained raters with those emerging from ratings by trained raters. Rating scales included in the study were two forms of behavioral expectation scales (BES), two forms of mixed standard scales (MSS), and graphic scales (GRS). A total of 206 undergraduate students evaluated one of two general psychology instructors using one of the five rating scales.

The results indicated that, for BES and GRS, training did not significantly improve the psychometric quality of the ratings. For MSS, results concerning the effectiveness of training were inconclusive. The results further revealed no psychometric differences among the rating scale formats when trained raters performed the ratings. Implications of the findings are discussed, along with recommendations for future research.