THE EFFECTS OF TRAINING AND TYPE OF ITEM ON INTERRATER AGREEMENT AND LENIENCY ERROR

By

Mark Douglas Spool

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Psychology

1978

ABSTRACT

THE EFFECTS OF TRAINING AND TYPE OF ITEM ON INTERRATER AGREEMENT AND LENIENCY ERROR

By

Mark Douglas Spool

Recently there has been a trend in research toward training raters to evaluate performance. Training programs designed to increase interrater agreement and/or reduce rating errors suffer in at least one of four major areas: (a) unit of analysis, (b) applied principles of learning, (c) focus on observation of behavior and (d) consideration of the difficulty of the rating task. The purpose of this dissertation is to answer two research questions: (a) "Would a training program that focuses on observation skills (i.e., observing specific behaviors) increase interrater agreement and reduce leniency error more than a training program that focuses on rating errors--given that both training programs apply some of the principles of learning and use the individual as the unit of analysis (by not having discussion among trainees)?" (b) "Do the results of training, whether it focuses on rating errors or on observation skills, depend upon the type of item being rated?"

Undergraduate students (n=168) from two psychology classes were randomly assigned to one of three treatment groups: (a) training directed toward developing their observation skills (Observation Training), (b) training, typical of those currently in use, directed

toward having them recognize rating errors (Rater Error Training) and (c) no training (Control). The interrater agreement and leniency error of instructor ratings by these groups were compared over two time periods and across four types of items: (a) specific-descriptive, (b) specific-evaluative, (c) general-descriptive and (d) general-evaluative. On a specific-descriptive item, the rater reports the occurrence of a specific behavior. On a specific-evaluative item, the rater makes a judgment about the quality of a specific behavior. On a general-descriptive item, the rater reports the occurrence of a more general, abstract "behavior." On a general-evaluative item, the rater makes a judgment about the quality of a more general, abstract "behavior" which must be inferred.

The Observation Training program, one hour and fifteen minutes long, focused on specific behaviors related to three general instructor behaviors. Trainees were shown examples of related specific behaviors. Trainees were told to base ratings of general behaviors on observations of the relevant specific behaviors. The Rater Error Training program, one hour long, focused on rating errors. Trainees were shown and practiced recognizing different rating errors. Trainees were told to avoid rating errors by rating the instructor for what he actually did. Trainees in both training programs independently practiced rating instructors in five short vignettes and received feedback with regard to what a group of experienced raters (judges) gave as ratings. There was no discussion in either group. One week after training, and again in four weeks, students in the three groups rated their instructor.

Analyses of variance (Treatment Group x Class x Time Period x Type of Item) revealed that only the Rater Error Training program was

effective in increasing interrater agreement, and this occurred only with general-evaluative items. The effects of Rater Error Training were consistent over both time periods and across classes. Neither training program was effective in reducing leniency error. Inspection of the average rating for the Control Group (M=3.10, with 1 being favorable and 5 being unfavorable), however, suggests that leniency error was not a problem to begin with.

The implications of these results call into question the practical significance of training raters, unless interrater agreement is of concern and the rating form contains general-evaluative items. Recommendations for future research include: (a) improving behavior observation training, (b) examining further the practical significance of rater training and (c) improving the measurement of rating difficulty vis-a-vis types of items.

DEDICATION

To Toushie

ACKNOWLEDGMENTS

I should like to express my gratefulness to my doctoral committee for their resourcefulness and encouragement in conducting and writing this dissertation. Dr. John Fry, who directed my dissertation, helped me in the development of the training programs. Dr. Neal Schmitt, my chairperson, provided advice in the statistical and methodological portions of the dissertation. Dr. Fred Wickert aided in the conceptualization of the study and provided much advice in the beginning. Dr. LeRoy Olson gave much of his time and knowledge in the development of the student rating of instruction forms as well as being the actor in the videotaped model of the instructor behaviors.

In addition to my committee members, I owe much to Dr. Steven Sachs, who contributed a lot of constructive advice in the writing of the training program scripts and videotaping, and to Jon Werner, who assisted me in the training programs and data collection.

I am also grateful for my family's support. My parents, Irene and Milton, and my brothers, Phil and Richard, all gave me encouragement to seek higher education. For this I thank them.

I am especially grateful to my wife Monda for her love, support, devotion and assistance (she videotaped my training programs), and to our son Aaron for his consideration in waiting until the training programs were conducted before being born.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Research on improving the psychometric quality of performance ratings has historically focused on the rating form (Weick, 1968). The development of forced-choice scales and behaviorally anchored rating scales are examples of some of the research efforts. However, because such efforts have not proven particularly fruitful (e.g., Borman & Dunnette, 1975), research recently shifted its focus to training raters to evaluate performance (e.g., Bernardin, 1978; Bernardin & Walter, 1977; Borman, 1975; Latham, Wexley, & Pursell, 1975).

Training programs in this area are designed to increase inter-rater agreement and/or reduce rating errors (e.g., leniency error) with the expectation of only achieving an improvement. For example, trainees in Bernardin's (1978) and Bernardin and Walter's (1977) training programs received a one-hour presentation on different types of rating errors and practiced recognizing them. Borman (1975) provided raters with a five to six minute presentation on halo error, including a definition and data illustrating it. Latham, Wexley, and Pursell (1975) compared a workshop with a group discussion approach in which raters discussed four types of rating errors. The studies evaluating these training programs, however, suffer from at least one of four major limitations which, in turn, may affect their results.

The first limitation is methodological. The training programs cited all require discussion among trainees about ratings and rating errors. When trainees interact with each other to the extent that each influences the other, as through discussions, then the unit of analysis should be raised from the individual level to the level of the group (i.e., trained group versus untrained group) for the assumption of independence among elements to be met. The unit of analysis is determined by the smallest level of data source (e.g., subjects or group of subjects). In each of the studies cited, however, the individual was the unit of analysis rather than the group. Statistical significance obtained by these studies, therefore, may be questionable.

The second limitation concerns principles of learning. Only the Latham et al. study applied at least some of the major principles of learning. The other training programs did not apply such basic principles as practice and immediate feedback (Spool, 1978). The effect of this limitation on training raters to evaluate performance is revealed in Latham et al.'s results. A training program (workshop) which provided trainees with practice, immediate feedback and realistic stimuli was shown to be more effective than a training program (group discussion) which did not apply some of the basic principles of learning.

The third limitation concerns the content of the training programs. These programs trained raters to recognize and, generally, "avoid" different types of rating errors (e.g., leniency and halo), which were presented either by lecture and/or by graphic illustrations and then discussed (Spool, 1978). A recent study by Borman (1978), however, revealed a need to refocus training objectives from increasing knowledge of rating errors and ability in recognizing them to increasing

skill in observing behaviors. In his study, a nearly ideal environment for obtaining error-free performance evaluations was created, part of which involved raters being very knowledgeable about various rating errors. Raters were even given a "refresher course" on rating errors. Still, interrater agreement was found to be far from perfect. One of the possible reasons for this finding, Borman hypothesized, was that the raters did not receive any training in observation skills. Thus, training raters to observe and recognize behaviors may prove more effective than training them to know or recognize rating errors.

The last limitation concerns the interaction of training with the difficulty of the rating task, as defined by the type of item in the rating form. Specifically, items in a rating form may be easy or difficult to rate, depending upon the level of the behavior (specific or general) and the amount of judgment (descriptive or evaluative) in the item (Borg & Gall, 1971). For example, if a rater were to rate an individual as to whether or not he performed a specific behavior, this rating task would be easier for him than if he were to evaluate a general, more abstract "behavior." Accordingly, training may be necessary for some items and not for others. Of the studies cited, none took into account the difficulty of the rater's task when evaluating the effectiveness of the training program. Training may appear to be effective or ineffective overall, but if the type of item (i.e., the difficulty of the rating task) were taken into account, different results (i.e., amount of interrater agreement or leniency error) may be obtained for different types of items.

These limitations raise two important questions. First, would a training program that focuses on observation skills increase interrater

agreement and reduce leniency error more than a training program that focuses on rating errors--given that both training programs apply some of the principles of learning and use the individual as the unit of analysis (by not having discussion among trainees)? And, second, do the results of training, whether it focuses on rating errors or on observation skills, depend upon the type of item being rated?

The purpose of this dissertation, therefore, is to answer these research questions by comparing the interrater agreement and leniency error, across types of items, of individuals receiving: (a) training directed toward developing their observation skills, (b) training, typical of those currently in use, directed toward having them recognize rating errors and (c) no training. In addition, these results will be assessed over time and across class settings/instructors. Specifically, the following null hypotheses will be tested: (a) the three treatment groups (a, b and c above) will not differ from each other across types of items on interrater agreement and leniency error, (b) the three treatment groups will not differ from each other across two time periods on interrater agreement and leniency error, and (c) the three treatment groups will not differ from each other across two classes on interrater agreement and leniency error.

Overview

In Chapter II, the pertinent literature on measures of reliability of ratings, reliability of student ratings of instruction, sources of rating errors, effects of level of behavior, judgment and content on ratings, principles of training, and training programs in performance evaluation are reviewed in detail. The design and

procedures of the study are discussed in Chapter III, and the results of the three treatment groups concerning interrater agreement and leniency error are presented in Chapter IV. Discussion of the results and conclusions appear in Chapter V along with comments for future research.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

The literature reviewed in this chapter appears under the following headings: "Measures of Reliability of Ratings," "Reliability of Student Ratings of Instruction," "Sources of Rating Error," "Effects of Specificity of Behavior, Judgment and Content on Ratings," "Principles of Training," and "Training Programs in Performance Evaluation." The Summary section which follows presents the generalities derived from the studies and literature.

The first three sections reviewed in this chapter provide the groundwork for the development and selection of the dependent variable measures in this study. The studies on the effects of specificity of behavior, judgment and content provide the information necessary for the development of the rating instrument, the content of the training programs and the design and analysis of the study. Finally, the information on the principles of training and the studies on training programs in performance evaluation serves as the basis for the development of the training programs. While the major part of the literature on training programs does not deal specifically with student ratings of instruction, some generalities may emerge from these studies which might be expected to hold true in the student rating situation.

Measures of Reliability of Ratings

Agreement among observers has most often been associated with reliability of measurement (Weick, 1968). In the area of behavioral observation, the term "consistency of observations" is used, and for performance evaluation it is "consistency of ratings." Whatever the term may be, "the conventions usually used to describe inter-observer agreement are far from complete and often deceptive" (Costello, 1973, p. 105).

Reliability has been estimated by several different indices and each may yield different results (Medley & Mitzel, 1963). The different measures of reliability are reviewed here and their strengths and weaknesses discussed with regard to their appropriateness as an indicator of interrater agreement.

One reliability estimate often used is a measure of internal consistency of the questionnaire. This measure represents the degree of similarity (i.e., homogeneity) of the items. It has been argued that internal consistency is an inappropriate measure for observer agreement. It primarily measures the degree of agreement between items, not raters (Kane, Gillmore, & Crooks, 1976), it is redundant when independent but simultaneous recorded observations are available (Costello, 1973), and it is a serious overestimate as a reliability estimate (Kane et al., 1976).

Three other common types or indices of reliability have been described by Medley and Mitzel (1963). The first index is the stability coefficient. It is measured by a correlation between scores based on observations made by the same observer at different times.

The second type of index is termed reliability coefficient. This term is used to refer to the correlation obtained between scores based on observations made by different observers at different times. Coefficient of observer agreement, the third index, is the correlation between scores based on observations made by different observers at the same time. These indices, according to Medley and Mitzel, may yield different estimates of reliability depending on the type of index that is computed.

The appropriateness of the stability coefficient has been questioned. It does not estimate the accuracy of a score since it is based on a correlation between observations made by a single observer. The "true" score, as Medley and Mitzel point out, pertains to the actual behavior which occurs, rather than to what some particular observer would see. A less powerful argument has been presented by Costello (1973). The coefficient of stability only tells us something about the consistency of behavior observed from time to time. Such data, Costello contends, are rarely questioned and usually do not concern us.

The reliability coefficient has been receiving increasing attention, particularly in the area of performance ratings. According to Medley and Mitzel, "only [it] tells us how accurate our measurements are" (1963, p. 254). As previously stated, the reliability coefficient is the correlation between scores from different observers at different times. Byrne (1964), however, cautioned that high interscorer correlations may be unreliable because it is possible for scorers to disagree on many items and yet have equal total scores; and it is possible

for one scorer consistently to give higher scores than the other, a difference that could not be detected in a correlation coefficient.

Most methods of estimating rater reliability, however, are analysis of variance procedures, predominantly the intraclass correlation coefficient (Showers, 1973). Ebel (1951) compared three formulas applicable to rating situations--average intercorrelation, the intraclass formula and the generalized formula for the reliability of averages--and concluded that the intraclass correlation formula is most versatile, allowing one to include or exclude "between raters" variance from the error term. For example, in the case of student ratings of instructors one would include between-raters variance in the error term because all raters do not rate all instructors. Also, as Engelhart (1959) points out, both a single rater estimate and an n-rater estimate can be obtained with the intraclass coefficient while the generalized formula only gives the n-rater case.

Cronbach, Rajaratnam, and Gleser (1963), as well as Kane, Gillmore, and Crooks (1976), suggest that the interclass formula allows one to generalize from randomly selected samples of raters to the reliability of raters in general. This is particularly desirable in determining the reliability of student ratings of instruction, since the particular group of students who were rating each instructor may more than likely be different every time.

There is a particular problem with the intraclass correlation coefficient, however, in that it "suffers from the limitation of requiring non-zero variance between items being rated in order to obtain a significant coefficient" (Finn, 1970, p. 71). Finn continues,

> It obviously follows that the number of items must be greater than one in order to apply the intraclass correlation. As a practical matter, one usually finds that the variance between items must be substantial in order to obtain an acceptable coefficient of reliability (p. 71).

The intraclass correlation coefficient, therefore, does not appear to be sensitive to the variance within items, an indication of the degree of agreement among raters.

An alternative method to estimating the reliability with which a group of raters rate items was proposed by Finn (1970). Finn's reliability index, represented only as r, does not make the requirement of non-zero variance between items and may be applied with any number of items. It is determined by subtracting from one the ratio of the variance obtained among items for all raters to the variance expected from random rating (no agreement). The variance of these expected random ratings where the distribution of possible ratings on a five-point scale is rectangular is 2.0. The degree to which the observed variance is less than 2.0 then reflects the presence of something other than error variance in the ratings. The ratio of the observed variance to the expected variance gives the proportion of random or error variance in the observed ratings. Subtracting this proportion from 1.0 then gives the proportion of non-error variance in the ratings, a reliability coefficient:

$$r = 1.0 - \frac{\text{Variance (observed)}}{\text{Variance (expected-random)}}$$

Finn compared this measure which primarily emphasizes the within-item variance, with the intraclass formula on several sample data. When degree of agreement among raters is the key focus, the

intraclass correlation is not suitable. It was found that if different items were rated similarly by the same rater (i.e., little between-item variance) and if each item were rated similarly across raters (i.e., little within-item variance or high degree of agreement among raters), the intraclass correlation was substantially lower than Finn's reliability estimate. When differences in ratings between items as well as between raters (low degree of agreement among raters) were increased, the intraclass correlation was substantially higher than Finn's. In both cases, Finn's estimate more adequately reflected the within-item variance, that is, the degree of agreement among observers.

The last measure of reliability of ratings treated here is observer agreement. Costello (1973) described three functions of such a measure assuming independence between observers: it shows (a) that the technique of observation is objective, and that personal bias could not produce the obtained results, (b) that the technique is communicable to others, and (c) that error variance from this source is low. Observer agreement is perhaps the most common reliability measure in observational studies (Weick, 1968). Kaplan (1964) argued that intra-subjectivity is preferable to replicability as a criterion, especially when rare events are observed.

One means of calculating observer agreement which has been widely used to describe the extent of differences between observers is simple-percentage agreement or more appropriately "percent exact agreement," i.e.,

$$\frac{100 \text{ (no. of agreements)}}{\text{(no. of agreements) + (no. of disagreements)}}$$

Using percentage agreement analyses can lead to misinterpretation, however. Raters may not agree exactly, but they may be close. According to the above formula, unless raters are in complete agreement, the ratio will be low or zero. A miss is as good as a mile.

Measures more adequately reflecting the degree of agreement among observers are those which represent the variance of ratings. Standard deviation of ratings on items has been used as a dependent variable in studies exploring interrater reliability of ratings (e.g., Bernardin & Walter, 1977). In these studies differences between items were ignored. In Bernardin and Walter's study, interrater reliability was derived by taking the standard deviation of the [three] ratings on each dimension for each ratee. The standard deviations were then taken as data points in a analysis of variance.

At least two problems exist with measuring interrater agreement in the above manner. First, their dependent measure represents the variance between items on a dimension, not the variance between raters. It does not, therefore, adequately reflect the degree of agreement among raters (i.e., interrater agreement); it is more of a measure of degree of agreement among items (i.e., interitem agreement). Second, differences between items were ignored. If different types of items can differentially affect raters, as evidenced by Spool (Note 1), then the measure Bernardin and Walter used is further complicated.

Perhaps a better approach would be to consider the differences between raters within each item and average these differences across similar items--similar in "type" (see Effects of behavior specificity, judgment and content of items on ratings section in this chapter).

This can be accomplished, following the standard deviation/variance concept, by taking the absolute value of the deviation of a rater's rating on an item from the average rating of that item for all raters, then averaging the absolute deviations across items within a "type." Such a measure would more adequately reflect interrater agreement--the degree of agreement among raters.

In summary, various measures of rater consistency or reliability have been reviewed. The strengths and weaknesses of each were discussed. Other measures (coefficients) of observer agreement not directly related to the present dissertation (e.g., Scott's coefficient, Cohen's kappa, Light's extension of K, Flander's modification of π, and Garrett's modification of π) are critically reviewed in Frick and Semmel (1978). It appears, as Medly and Mitzel (1963) have warned, that different indices may yield different estimates of reliability; hence, selection of an appropriate measure is extremely important. The measure adopted as the dependent variable in this study is discussed in the Methods section and takes into account the problems noted above.

Reliability of Student Ratings of Instruction

There have been three major reviews on the reliability of student ratings of instruction: Costin, Greenough, and Menges (1971), Kulik and McKeachie (1975) and Doyle (1976). In these reviews, the two reliability estimates most often used have been measures of stability (consistency over time) and homogeneity (internal consistency). None of these review articles mentioned the use of other measures of reliability, like intraclass correlation. At least one study, however, exists which used the intraclass correlation coefficient (Showers, 1973).

The general findings from studies on internal consistency and stability have been consistent. Costin et al. report stability correlations from .48 to .89 between student ratings of instructors repeated at intervals from two weeks to one year. Internal consistency coefficients are typically high--in the .80s and .90s (Kulik & McKeachie, 1975). "It would appear, then, that students can rate classroom instruction with a reasonable degree of reliability [stability or internal consistency]" (Costin et al., 1971).

Doyle (1976), however, questioned the generality of this conclusion and called for a consideration of the various purposes of evaluation. For example, for course improvement purposes the level of internal consistency and stability coefficients appear at an adequate level and also seem reasonably free from random error. However, for personnel decisions, the reliability of measures must be greater. Both random error and systematic error need to be reduced. "The conclusion [then] is not to avoid using student ratings in personnel decisions but rather to take steps to improve the precision of these measures" (Doyle, 1976, p. 44).

It should be noted that none of the studies reviewed considered the degree of agreement among students, a kind of reliability estimate for which there is certainly a logical appeal. For example, we might expect an instructor to place more credence in student feedback when the students are in agreement with one another. More specifically, items on a rating form in which students rate with greater agreement should have greater impact upon the instructor than items in which there is less agreement. Neither measures of stability nor measures

of internal consistency adequately reflect this kind of degree of agreement. It is possible that students may not be in agreement with each other even though students rate the same over time (high stability) or the items in the form are highly related to each other (high internal consistency). Therefore, a measure of interrater agreement to evaluate student ratings of instruction is warranted.

Sources of Rating Error

Errors in making subjective judgments or ratings stem from several sources (e.g., raters' personal tendencies, contamination from extraneous sources and inappropriate weighting of factors) (Dunnette, 1966; McCormick & Tiffin, 1974). The most pervasive source of error in performance ratings by a rater-observer are response tendencies (also called biases or sets). Response tendencies exist when the rater-observer completes all rating forms in about the same way for all ratees, failing to discriminate either among different persons or within the behavioral repertoire of a single person, irrespective of the actual behavior of the ratee.

Three rating errors representing the most common response tendencies are:

1. Central tendency. Ratings by a rater-observer who commits the central tendency error are characterized by scores averaging in the midpoint of the rating scale with a small standard deviation (e.g., less than 1.00).
2. Leniency. Here, the rater-observer says only "good" things about everyone. Ratings by one who commits the leniency error are characterized by scores falling within the top two points (i.e., the favorable end) on the rating scale.

3. Halo. The observer-rater, in completing the rating form, may make an overall evaluative judgment about each ratee and then proceed to describe him or her with all "good"-sounding or "bad"-sounding ratings, irrespective of the actual behavioral content of the items in the rating form.

In the area of student ratings of instruction, leniency has been identified as the major source of rating error; student ratings of instruction seem to be overwhelmingly concentrated at the upper end of the rating scale when there is reason to believe that they should not (Showers, 1973). Studies measuring leniency error simply examine the closeness of the ratings to the midpoint of the rating scale.

Effects of Behavior Specificity, Judgment and Content of Items on Ratings

In the area of behavior observation and performance ratings, there appears to be three major classes of variables which may have an impact on the reliability of ratings: the specificity-generality of behavior observed, the degree of judgment required (i.e., whether observers are required to describe/report or evaluate the behavior), and the content area of an item. Each of these classes of variables will be discussed, including research results as well as general conclusions. It should be noted here that there have been more general statements made about these classes of variables than there has been research. Gellert succinctly described the situation back in 1955, and not much has changed since then. "Although the difficulty of reliable recording has been recognized from the beginning, little systematic research has been done with regard to factors that affect

the reliability of different observation methods when they are used by various observers" (p. 186).

Behavior Specificity

The degree of specificity-generality of the behavior to be observed and recorded influences the level of inference required of an observer. The term "inference" refers to the process intervening between the objective behavior seen or heard and the coding or rating of this behavior on an observational instrument (Rosenshine, 1971). Unfortunately, there does not exist an index measuring this variable on a continuous scale. At best we can speak of specific behaviors/items and general or global behaviors/items (Heyns & Lippitt, 1954; Rosenshine & Furst, 1971). Specific behavior items "focus upon specific, denotable, relatively objective behaviors such as 'teacher repeats student ideas,' or 'teacher asks evaluative questions'" (Rosenshine & Furst, 1971, p. 19). General behavior items are named such because they lack the concreteness of specific behavior items. "Items on rating instruments such as 'clarity of presentation,' 'enthusiasm,' or 'helpful towards students' require that an observer infer these constructs from a series of events. In addition, an observer must infer the frequency of such behavior in order to record . . . it . . . somewhere on the set of gradations used in the scale of the observational rating instrument" (Rosenshine & Furst, 1971, p. 19).

Research on the effects of specificity-generality of behavior is minimal. In general, it has been found that discrete (behaviorally specific) items seem to yield higher levels of observer agreement

than items which require agreement about more global behavior, such as traits (Becker, 1960; Walter & Gilmore, 1973).

General statements about this variable, on the other hand, are in abundance. Gellert (1955) attributed poor reliability to two sources: faulty research design (such as inadequate training of observers) and inconsistency among observers. The latter source can be a result of using very general behavior items. More specifically, "the more inference is involved in making judgments [ratings], the greater the opportunity for disagreement between judges" (p. 192).

Several recommendations have been made in regard to the use of general behavior items. Most observation manuals, for example, suggest the use of specific behavior items only, or at least the minimization of the general behavior items (e.g., Medley & Mitzel, 1963; Weick, 1968). Other researchers highly recommend the use of both specific behavior items and general behavior items (e.g., Cartwright & Cartwright, 1974; Heyns & Lippitt, 1954; Gellert, 1955). Still others who would permit the use of general behavior items add certain stipulations. Borg and Gall (1971) qualify their use of general behavior items. They recommend if such items are used that the researcher provide the observer with several examples of each variable. Rosenshine (1971) suggests the use of specific behavior items as cues, preceding a related general behavior item, to help in the rating decision. Training of observers has also been a stipulation if general behavior items are to be used (e.g., Borg & Gall, 1971; Medley & Mitzel, 1963). Long and Gall recommend the following: conditions for training:

> Ideally the researcher might videotape instances of behavior illustrating points of the continuum that define the variable (e.g., videotape teachers who are very confident, teachers who are somewhat sure of themselves, and teachers who are confused and anxious). The observer would study [i.e., be trained on] these examples before attempting to record observations for the main study. (1971, p. 227)

Judgment

Degree of judgment can be defined as whether the observer reports or describes behavior or whether the observer evaluates behavior. Evaluative variables are related to behavior specificity variables since they also require an inference from behavior on the part of the observer (Borg & Gall, 1971). However, they have an additional requirement in that they refer to the quality (a value judgment) of the behavior.

Statements about the degree of judgment as a variable impacting upon the reliability of ratings have generally been in favor of descriptive rather than evaluative items. Descriptive responses require a minimum of inference (Jones, Reid, & Patterson, 1975). On the other hand, the "reliability of evaluative measures requires agreement on both the topography and the intent of a behavior (Weinrott, 1975, p. 13). The general recommendation to get high reliability, therefore, is to either avoid using evaluative items or to at least provide examples of behavior that define points along the continuum of excellent-to-poor explanations (Borg & Gall, 1971). Such efforts have been manifested in the work on behaviorally anchored ratings scales, BARS, or behavior expectation scales, BES (cf. Smith & Kendall, 1963).

Research, however, has not tended to support the fruitfulness of these recommendations. In a recent study, Borman and Dunnette (1975) compared behaviorally anchored scales to scales containing the same dimensions and definitions but without behavioral anchors and to a series of scales involving trait-oriented dimensions, also without anchors. Their findings showed that the magnitudes of the differences due to formats were small, in no case exceeding more than five percent of the variance on the dependent variable. In other words, even though statistically significant differences were found, there were no practical differences. Their recommendations, if there were to be a continuation of the use of BARS, was that raters should be trained. Finn (1972) and Showers (1973) also obtained similar results. In Finn's study, he additionally compared items with a descriptive scale to a simple numerical scale and found no differences.

Content

It appears that no research has been conducted where content of the item was studied to determine its influence on the reliability of ratings. Only one study reported results which have a bearing on this area (Showers, 1973). In her study, different formats (i.e., response cues) of a student rating of instruction form were systematically compared with leniency bias and interrater reliability (measured by the intraclass correlation coefficient) as dependent variables. Although it was not intended as the focus or purpose of her study, Showers found differences in rater reliability according to item content. Specifically, Student-Instructor Interaction items were found to be more reliably rated than items related to Student Interest and Course

Organization. This finding seems plausible if one considers the differential impact instructor behaviors (i.e., items) may have on students. Items related to instructor behaviors which directly affect students' learning of the subject matter (e.g., stating the course objectives or asking students questions) may be observed more accurately and recalled better, so that students would be in greater agreement with each other, than items related to instructor behaviors which may have less of an impact on the students' learning. Items of the latter type include behaviors related to motivation or class structure. The instructor behavior "being interested in the subject matter," for example, may not be noticed/observed by many students at all. Consequently, such items should have less interrater reliability.

Principles of Training

To obtain a better understanding of the problems of the studies to be reviewed and the reasons behind the development of the training programs used in this dissertation, a brief review of the principles of learning and motivation and the principles of transfer of learning are in order. These principles collectively will be referred to as principles of training. In this section, the principles of training will be presented and briefly described. Following will be a discussion of the principles of transfer of learning. The possible application of these principles to training will be included in the discussion.

A comment is appropriate before proceeding with a description of the principles of training. The bulk of original research from which the principles of learning and motivation were developed was conducted on animal and not human subjects. From there, laboratory

studies using humans involving simple to complex tasks were conducted. The tasks often involved serial order or word association learning of nonsense syllables. The application of research results from studies such as these to the training setting is still in an exploratory stage with no firm answers available. The best that can be done is to take what has been learned in the laboratory and suggest what can be done during training.

A summarization of the eight principles of learning and motivation relevant to training can be found in Davis, Alexander and Yelon (1974):

1. Meaningfulness. A trainee is more likely to be motivated to learn things if they are meaningful to him. For subject matter to be meaningful, a trainee should be able to relate to it personally. This can be accomplished by relating the training material to:
   (a) trainees' past or present (e.g., job) experiences
   (b) trainees' interests and values
   (c) trainees' future activities or aspirations
2. Prerequisites. A trainee is more likely to learn something new if he/she has all the prerequisites. Having the prerequisites will more than likely make the training meaningful; trainees will be capable of perceiving relationships between relatively simple knowledge they possess and more complex knowledge that they are asked to learn.
3. Model. A trainee is more likely to learn if he/she is presented with a model of the behavior to be learned.

Therefore, trainees should be presented with a model, the strategy or plan of attack beforehand; all steps should be pointed out and labeled; an explanation of why all decisions made should be given and all consequences pointed out.

4. Open communication. Learning is facilitated if the training is structured and the trainer's messages are open for inspection. To accomplish this, trainees should be told what they are to learn, how well they are to learn it, and under what conditions they are expected to perform the rating (i.e., learning objectives). Also, they should be told why the task is important and how learning to rate fits into a bigger picture.
5. Novelty. If trainees' attention is attracted by relatively novel presentations, they are more likely to learn. Therefore, various modes of presenting the same material should be used during the training, as well as variety, such as modulating the presenter's voice and using several different examples for each point.
6. Active appropriate practice. No learning occurs without practice. However, practice must be relevant to what is to be learned. Therefore, trainees should be required to practice rating in realistic approximations of on-the-job situations, answer questions by writing them down, give examples, etc. Further, practice should be scheduled in short periods distributed over the time allocated for training.
7. Fade prompts gradually. At the beginning of training, trainees should be provided with prompts or hints (cues). As they

become proficient, the prompts should be systematically withdrawn or faded out.

8. Pleasant conditions and consequences. A trainee is more likely to continue learning if instructional/training conditions are made pleasant. To provide pleasant conditions and consequences, the negative aspects of instruction should be eliminated and the positive should be accentuated. Instructions should be written so as to avoid unpleasant physical conditions. Also, challenging tasks should be set and trainees given feedback as to their practice as soon as possible.

While these eight principles of learning and motivation presented above are listed separately, in practice they are interactive. Moreover, these principles do not cover all important aspects of learning; principles related to transfer of training are also important. The principles of learning and motivation are concerned with the acquisition of knowledge while the principles of transfer of training are more concerned with retention and use in new settings or novel situations.

Four basic principles of transfer, summarized from the research literature in transfer of learning (Goldstein & Sorcher, 1974), should be applied to the design of any training program:

1. General principles. Transfer of training is facilitated if the trainee is provided with general, mediating principles governing satisfactory performance on both the training and beyond-training setting. This general principle should include stating the consequence of behaving or not behaving as presented in the training session: the reasons why. It also includes the provision of learning sets, such as "advance organizers."

2. Response availability. Transfer of training has also been demonstrated to be facilitated by procedures that maximize response availability. Therefore, training should be organized so that easier tasks are attempted first (a higher likelihood of correct responding of the trainee) with immediate feedback (reinforcement). Subsequent tasks should be made increasingly more difficult and should be followed by partial or intermittent reinforcement.
3. Identical elements. The greater the number of identical elements or characteristics between the training and application settings, the greater the consequent transfer. This is perhaps the most neglected principle in rater training. Therefore, trainees should practice rating the performance of people in settings which simulate real conditions, with fidelity as high as possible.
4. Performance feedback. If what is learned during training is to endure beyond the training setting, then feedback on the behavior or performance of the trainee must continue. Otherwise, extinction (i.e., complete removal of reinforcement resulting in non-performance of desired behavior) will occur. Therefore, the training program should make provision for assisting the trainee in self-reinforcement after training.

In summary, there are several principles of training which are important to consider when developing a training program. However, as advice to the novice trainer, it may prove difficult to implement all of these principles in any given training program, but their

combined impact is to greatly increase the likelihood of satisfactory learning and positive transfer (Goldstein & Sorcher, 1974).

Training Programs in Performance Evaluation

This section includes research directed at the examination of any technique which has been designed to improve observers' skills of observation prior to their entering an observation setting. This is in contrast to research on category or rating scales. Thus, only studies which investigated the effects of observer training on observer accuracy are included. Likewise, studies for which criteria for evaluation are unrelated to the problem of observer accuracy (e.g., communication skill) are not included in this review. Others, although they may involve a training program, are not included if the purpose of the study was not to improve observer accuracy, but instead was to determine the effects of other variables (e.g., order of observees) on observer accuracy. However, some studies in which the major emphasis was not on the evaluation of the observer training program are included because they described a "training program" to minimize the observer accuracy problem.

Six studies were found to be related to performance evaluation/rating. The studies are summarized in Table 1 and analyzed along the following dimensions:

1. Type of behavior observed (verbal and nonverbal)
2. Role of the observer (descriptive, inferential or evaluative)
3. Whether more than one training program was compared
4. Training design (i.e., components and length of training program)

Table 1

Summary and Analysis of Research on Training Programs: Reducing Rater Bias Evaluation

Study	Training context	Beh[a]	Role[b]	>1?[c]	Training design	Experimental design	Criteria of evaluation (level of criteria)	Results
Levine & Butler (1952)	performance ratings	NV (V?)	E	yes	three groups 1) discussion 2) lecture 3) control	Pretest/Post-test Control Group	halo effect (behavior)	group: 1) effective in modifying supervisors' rating behavior (i.e., reducing rating error) 2) no influence on changing the supervisors' methods of rating 3) no observable behavior change on the part of the supervisors (i.e., rating error not eliminated)
Borman (1975)	performance ratings	NV	E	no	5-6 minutes; instructions: definition of halo & presentation of data illustrating the error	One-Group Pre-test/Posttest	halo effect, inter-rater reliability and validity (behavior)	the short training session significantly reduced halo between ratees, not within ratees; interrater reliability was lower after training; validity was unchanged
Latham et al. (1975)	performance appraisal and selection interview	V & NV	E	yes	three groups: 1) workshop (modeling through videotape, practice, discussion and feedback 2) group discussion 3) control	Posttest-Only Control Group	four rating errors: 1) similarity 2) first impressions 3) contrast effects 4) halo effect (behavior)	group: 1) trainees committed none of the errors 2) trainees committed impression errors 3) committed similarity, contrast and halo errors
Burnaska (1976)	performance rating	V & NV	E	no	presented and discussed rating scales and common measurement errors; behaviors displayed on videotapes and discussed ratings compared to criterion ratings	none (not evaluated)	interrater reliability (behavior)	training effective ("gut" reaction)
Bernardin & Walter (1977)	student ratings of instruction	V & NV	D, I & E	yes	four groups: 1) 1 hr. total: presentation on rating errors; reviewed BES; raters kept observation diary on their instructor (training occurred during 1st week of semester) 2) same as Group 1 but BES not seen (only discussed) and no diary kept 3) same as Group 2 except occurred immediately prior to rating 4) control group (received written instructions only about rating errors)	Posttest-only Control Group	leniency effect; halo effect; interrater reliability; discrimination across ratees (behavior)	Group 1, which received psychometric training and exposure to the evaluation scale prior to and during observation, showed significantly less leniency error and halo effect than all other groups; interrater reliability was also higher for Group 1 than for Group 2; there were no significant differences between groups in assessing discrimination across ratees

Table 1 (con't)

Study	Training context	Beh[a]	Role[b]	>1?[c]	Training design	Experimental design	Criteria of evaluation (level of criteria)	Results
Bernardin (1978)	student ratings of instruction	V & NV	D, I & E	yes	four groups: 1) 1 hr. involving definitions, graphic illustrations & examples of rating errors; practice recognizing errors with discussion (cf. Bernardin & Walter, 1977) 2) 5 min. lecture/presentation on rating errors (cf. Borman, 1975) 3&4) control groups (they were told of the importance or unimportance of ratings, respectively)	Pretest/Posttest Control Group	internal (knowledge of rating errors) & external (errors committed in rating) criteria: leniency error (i.e., shift in mean ratings from the midpoint of the scale in the favorable direction); halo effect (the standard deviation compiled across dimensions of a rater's several ratings of a particular ratee) (behavior)	Group 1 was superior to Group 2 in psychometric quality of ratings and both groups were superior to the control groups at the 1st measurement; no differences were found between any groups in later comparisons; a consistent relationship was found between the internal and external criteria

[a]Type of behavior observed: V = Verbal
NV = Nonverbal

[b]Role of observer: D = Descriptive
I = Inferential
E = Evaluative

[c]More than one training program compared?

5. Experimental design used to evaluate the training program (cf. Campbell and Stanley, 1963)
6. Criteria used for each evaluation (e.g., interrater reliability and halo effect)
7. Results of each evaluation

In conjunction with the criteria of evaluation, the level of criteria (cf. Kirkpatrick, 1959) is also assessed. Kirkpatrick suggests that four different levels of criteria should be used to determine the effectiveness of a training program: reaction, learning, job behavior and organizational results.

All of the studies involved the observer as an evaluator and, as a criterion measure, used either a measure of rater bias, interrater reliability or both. Four of the six studies compared more than one training program. Only these four will be discussed in detail. One of the remaining studies (Borman, 1975) had an extremely minimal training program (five minutes of instruction). In this study, trainees read about, rather than actually observed, subjects' performance. The other study (Burnaska, 1976) did not evaluate its training program (Burnaska, Note 2).

Levine and Butler (1952) dealt with supervisors who were over-rating employees in the higher job grades and underrating those in the lower grades--a form of halo effect. The supervisors were randomly assigned to a control, a lecture or a discussion group. In the lecture group, theory and technique of performance rating were presented. The lecturer also explained the problem caused by the previous ratings and what each supervisor needed to do to correct the problem. Questions

were encouraged and answered by the lecturer after his/her presentation. In the discussion group, the supervisors met together to discuss the nature of the problem and how it could be solved. The discussion leader merely acted as a moderator, avoiding interjection of his/her own opinions. After generating a number of ideas, the group arrived at one decision acceptable to all.

The results showed that no observable behavior change occurred on the part of the supervisors in the control group. Similarly, the lecture had no influence on reducing the halo effect. Only the group discussion method was effective in modifying the supervisors' rating behaviors (i.e., reducing halo error). Limitations of this study, however, were that it dealt with the elimination of only one rating error, the effects of the training were not assessed over time, and the unit of analysis should have been the group, not the individual, because the treatment involved group discussion and interaction. Also, the trainees did not practice and receive feedback.

Another study, Latham, Wexley, and Pursell (1975), interpreted the findings of Levine and Butler as meaning that knowledge of rating errors alone will not lead raters to take effective steps to counteract them. Latham et al. hypothesized that only an intensive workshop, as per Wexley, Sanders, and Yukl, 1973, and described in Spool (1978), would be effective in reducing rater bias. They compared a workshop, a group discussion and a control group. In their study, managers in each of these groups eliminated rating errors that occurred in performance appraisal and selection interviews, namely, contrast effects, halo effects, similarity and first impressions. The workshop included

videotapes of hypothetical job candidates being appraised by a manager. The trainees gave a rating of how they thought the manager in the videotape evaluated the candidate. Group discussion followed as to the reasons for each trainee's rating of both the manager's evaluation and the candidate. Thus, trainees viewed a videotaped model, had an opportunity to practice and received feedback.

In the group discussion method, the definitions of the four rating errors were presented and an example of each error was given for three situations: the performance appraisal, the selection interview and an off-the-job situation. The trainees then generated and shared (discussed) with each other personal examples of rating problems. Solutions to these problems were then generated and shared. These solutions turned out to be identical to those decided upon in the workshop group. Thus, the primary difference between the two training programs was the method used to eliminate the rating errors.

Six months after training, the managers rated hypothetical candidates who were observed on videotape. The results showed that: (a) observations of trainees in the control group were characterized by similarity, contrast and halo errors, (b) trainees in the group discussion made impression errors, and (c) observations of trainees in the workshop were relatively free of all the errors. A possible limitation of this study is that the testing was a simulation rather than an actual measure of the trainees' on-the-job behavior. Furthermore, because the treatment groups incorporated discussion, the legitimacy of using the individual manager, rather than the group, as the unit of analysis must be called into question.

Bernardin and Walter (1977) investigated the effects of different training programs on rating errors of students evaluating faculty performance on behavior expectation scales (BES). Students (n = 156) of 13 different instructors were randomly assigned to one of four experimental groups. Group One received one hour of training in the first week of the semester on the various types of rating errors. Definitions and graphic illustrations/examples were given. For example, for halo error, a distribution of ratings were presented for three raters across seven dimensions of work performance. Students were asked to judge which raters were guilty of halo error. A general discussion of rating error followed. Students were also given copies of the seven BES to be used in the tenth week of the semester to evaluate the instructor. They were asked to maintain an observational diary for their instructor by recording observed critical incidents on the BES throughout the semester as they pertained to the seven dimensions of performance. It was pointed out that the diary would be collected and checked at the end of the semester. Although trainees in the group had the opportunity to practice, they did not receive feedback since the diary was collected and evaluated after the evaluation time. Group Two received identical training on the same types of rating errors as Group One, also in the first week of the semester. The BES was discussed but not seen. The students, however, were instructed to observe the instructor throughout the semester with the performance dimensions in mind. Group Three received the identical training on rating errors and participated in a similar discussion on the BES as did Groups One and Two. Their training and discussion, however, took place in the tenth week of the semester,

immediately prior to the evaluation of the instructors. Group Four received no training prior to the period of evaluation, but brief mention was made of halo error and learning effect in the written instructions of the evaluation form.

For each instructor of the course there were three raters from each of the experimental groups. The results revealed that Group One, which had received psychometric training and exposure to the evaluation scale prior to and during observation, showed significantly less leniency error and halo effect than all other groups. Interrater reliability, derived by taking the standard deviation of the three ratings on each dimension for each ratee, was also higher for Group One than for Group Two. Further, the two groups differed with respect to halo error: the greater the emphasis on observation, the less halo error there was. There were no significant differences, however, between groups in discriminating across ratees.

The differences between Groups One and Two illustrate the importance of familiarization with the rating scales. As Bernardin and Walter point out, the recommendation by Smith and Kendall (1963) to bring appraisal into closer correspondence with observation was thus applied to Group One.

This study suffers from several major problems. The first concerns the question of whether the subjects really received practice in rating. During the one hour of training, Group One subjects were shown the rating scales but did not use it to rate anything (e.g., videotapes). The only practice subjects received during this hour was in recognizing different psychometric rating errors (Bernardin, Note 3).

Although subjects kept a diary of observed critical incidents on the BES throughout the semester as they pertained to the seven dimensions of performance, it did not appear that the subjects actually rated their instructor on the rating scale. Related to the point about the diary, none of the subjects received feedback about the accuracy of their observations. Diaries were collected and checked at the end of the term but subjects were not told of their results. The authors stated that only three of the 39 members of Group One turned in diaries that were unquestionably of poor quality (less than three incidents per dimension--the average was 3.9--which were mostly of an ambiguous nature). The subjects, however, did not know this. And even if they did, there would have been no time for remediation. Feedback must be given as close to the desirable behavior as possible with opportunities for more practice and feedback (Goldstein & Sorcher, 1974). Further, the content of feedback must reflect the desired objective of training.

Other problems of a lesser magnitude include the issue of the unit of analysis. Again, as in the other studies, the treatment is highly characterized by discussion among subjects. This raises the unit of analysis to the level of the group. Analysis, however, was done at the individual level. Another problem which may affect the generalizability of the study is that subjects in Group One were exposed to the rating scale in the first week of classes. The other groups were not. It could be argued that Group One surpassed the other groups solely because these subjects "knew" what to look for and not because of any observation skills they might have developed from training. A better design would have subjects practice and

receive feedback on a similar but not identical (i.e., parallel) rating form as the one used to evaluate the effectivenss of observation training.

The last of the four systematic research studies in the area of reducing rater bias (Bernardin, 1978) compared a short lecture-training session with a more comprehensive (participative) training program, as per Bernardin and Walter (1977), in the context of student ratings of instruction. More specifically, the former training group, described by Borman (1975), received a five minute lecture on the definitions of three rating errors (leniency, halo effect and central tendency) with a presentation of one graphic illustration for each error. Reference was made to the seven performance dimensions subsequently measured on the rating instruments but students were not shown the actual scales. The latter training group (one hour in length) involved definitions, graphic illustrations, and examples of the same three errors. Students were also given data to evaluate in terms of the errors and the evaluations were discussed. Reference was also made to the dimensions on the rating instruments.

Eighty undergraduate students (20 in each of four groups) rated all of their instructors over one, two or three rating periods using behavioral expectation scales or summated rating scales. Tests on psychometric error (leniency and halo) were also administered at these times. This study, therefore, was able to explore issues not addressed by the other studies; specifically, stability of training and the relationship between internal (i.e., knowledge of psychometric error) and external (i.e., errors committed in performance ratings) criteria of rater training effects.

A consistent relationship was found between scores on the tests of psychometric error and error as measured on the ratings. Results also indicated that the psychometric quality of ratings was statistically (but not practically) superior for the group receiving the comprehensive training, and both training groups were superior to the two control groups at the first measurement period. The major finding of Bernardin's study, however, was the diminishing effect (non-stability) of training over time. No differences were found between any groups in later comparisons; the effects were virtually gone after one rating period. Bernardin, therefore, questioned the cost-benefit investment of the more comprehensive training program.

It should be noted, however, that the same problems as with the comprehensive training program criticized in the Bernardin and Walter (1977) study, such as trainees not having an opportunity to practice rating and not receiving feedback, exist. Perhaps, if the training program were improved (i.e., incorporating these essential principles of training that were missing), better results might be obtained.

In summary, the studies reviewed showed that accuracy in observation (i.e., interrater reliability and psychometric quality of ratings) can be statistically improved by training observers to avoid rating errors. However, most of the training programs lacked essential principles of training, such as active appropriate practice with feedback. Furthermore, the methodology used to evaluate training left much to be desired. The major problems were lack of a control group or use of an inappropriate unit of analysis. The practical success of the training programs, then, must be questioned, particularly if

improvement in observation accuracy is only slight. There may be a possibility of practical, as well as statistical, improvement if a training program were to incorporate the essential principles of training.

Summary

The review of the different measures of reliability revealed several problems with commonly used measures (e.g., internal consistency, reliability coefficient, percent agreement and intraclass correlation). Most of the problems center around the issue that the measure may not be adequately reflecting the degree of interrater agreement. Measures which represent the variance of ratings between raters may be more appropriate. One such measure was suggested.

The literature on the reliability of student ratings of instruction indicated that the two major measures have been internal consistency and stability. One study used the intraclass correlation, but none considered a measure of degree of agreement. An argument for its use was presented.

The review on sources of rating error revealed that among the several possible sources, student ratings of instruction appear to be most prone to leniency error. Here, the tendency among raters is to rate everyone toward the favorable end of the rating scale.

The review of the effects of behavior specificity, judgment and content of items on ratings showed little direct evidence in the area of student ratings of instruction. Indirect evidence, however, suggests that items which require raters to rate general behaviors are more likely to yield greater disagreement among raters than items containing

specific behaviors. Further, items which require raters to evaluate the behavior are likely to result in greater disagreement among raters than items in which the raters just report or describe the behavior. Also, content of the item may influence degree of interrater agreement. Items in which the content directly relates to the learning of students (e.g., "The instructor asks the students questions") should yield greater interrater agreement than items which indirectly relate to the students' learning (e.g., "The instructor maintained the attention of the students").

The principles of training (i.e., learning and motivation and transfer of training) review gave an indication of the various components of a training program which are essential to the learning of trainees. The eight most relevant principles of learning and motivation, followed by four basic principles of transfer of training, were defined and discussed in regard to their application to the development of a training program.

The review on training programs in performance evaluation covered six studies, of which only three were discussed in detail because of their possible contribution to the present study. Overall, training was found to statistically increase interrater reliability and improve the psychometric quality (i.e., reduce rater bias) of ratings. The objective of these training programs was to make trainees knowledgeable about and able to recognize rating errors. Problems related to (a) the content of the training programs, (b) the process of the training programs, and (c) the methodology of the studies evaluating the training programs were noted. Specifically, these

training programs (a) did not focus on observation skills, (b) lacked essential learning components, such as practice and feedback and ignored for the most part issues such as rating difficulty, and (c) used an inappropriate unit of analysis and/or lacked a control group when evaluating training. The practical success of the training programs, therefore, was called into question. Suggestions for the improvement of training programs in this area were discussed.

CHAPTER III

METHOD

Sample

Undergraduates (n = 168) from two psychology courses, described below, at a large Midwestern university participated in the present study. The subjects, primarily sophomores, junior and seniors, represented a variety of majors (e.g., psychology and business).

Design

The present study was a 3x2x2x4 factorial design--Treatment Group (G), Class (C), Time Period (T) and Type of Item (I)--with subjects nested within GxC and crossed with TxI (see Figure 1). Time Period and Type of Item were repeated measure variables.

Independent Variables

Treatment Group

Subjects were assigned to one of three groups, (a) Behavior Observation Training Program, (b) Rater Error Training Program, and (c) Control Group, according to the following procedure: students who volunteered for the present study were randomly assigned to one of the two training programs; subjects in the Control Group were randomly selected from the remaining students who came to class the day of the ratings.

		Time Period								
		T_1 Type of Item				T_2 Type of Item				
Treatment Group	Class	I_1	I_2	I_3	I_4	I_1	I_2	I_3	I_4	
G_1: Behavior Observation Training	C_1									$n_{11}^a=28$
	C_2									$n_{12}=28$
G_2: Rater Error Training	C_1									$n_{21}=28$
	C_2									$n_{22}=28$
G_3: Control Group	C_1									$n_{31}=28$
	C_2									$n_{32}=28$

$^a n_{ij}$, where i = G and j = C

Figure 1. Design of the Experiment

Both training programs were held on two consecutive evenings during the second to third week of the term. No class sessions for the classes involved were held between these evenings. To insure standardization across evenings, the training programs were videotaped presentations. Scripts for these training programs are found in Appendix B.

Subjects in the Behavior Observation Training Program were trained to observe instructors' performance of specific behaviors and to rate instructors on general behaviors related to "maintaining attention." Training emphasized how to rate general behaviors based upon observations of specific behaviors.

The Behavior Observation Training Program consisted of six parts: (1) Introduction, (2) Presentation on rating general behaviors, (3) Explanation of the rating scale, (4) Explanation and demonstration of specific behaviors related to three general instructor behaviors, (5) Practice observing and rating with feedback, and (6) Summary. The training lasted one hour and 15 minutes.

Part One: The introduction began with a videotaped presentation of a common situation--students rating the same instructor differently. Subjects were asked (rhetorically) "Why?" Then, the purpose, ground rules (e.g., no discussion) and overview of the training program were given.

Part Two: Following the introduction, a short videotaped vignette was presented depicting two students disagreeing about their rating of an instructor on a general behavior. The trainer then posed the following questions: "Why do you think there can be such a wide difference in (the two students') views?" "Is one right and the other

wrong, or can both be right?" An explanation of why differences in ratings of general behaviors exist was given. Then a procedure for subjects to follow when rating general behaviors was presented: (a) to observe those specific behaviors which are related to the general behavior to be rated; (b) to try to recognize situations when these specific behaviors could occur; that is, when opportunities for their use arise; (c) to consider the occurrences and/or non-occurrences of the specific behaviors; and then, (d) to rate the instructor on the general behavior (i.e., the extent to which the subject agrees that the general behavior is characteristic of the instructor). The following general principle was also given: "Base ratings of general behaviors on observations of the presence or absence of relevant, specific behaviors."

Part Three: The explanation of the rating scale included an example of the rating form used in rating the three general instructor behaviors on a Likert-type scale which ranged from Strongly Agree to Strongly Disagree. The general principle was reemphasized. The following instructions were given: "If the instructor does many of the specific behaviors related to the general behavior, then you should rate him or her Strongly Agree or Agree, depending upon the number of specific behaviors occurring. However, if the instructor does only a few or none of the relevant specific behaviors, then you should rate him or her toward the Disagree or Strongly Disagree end of the rating scale, again depending upon the number of specific behaviors occurring or not occurring."

Part Four: The three general behaviors ("made the subject relevant," "helped students keep their attention on the subject matter,"

and "was enthusiastic when lecturing") were described by the trainer, first with a general definition and then with a description of each relevant specific behavior. Videotaped vignettes were then presented of an actor (a Learning & Evaluation Service faculty member) demonstrating each general behavior. There were two vignettes for each general behavior. In the first vignette, all the specific behaviors were present; therefore, the instructor would be rated at the Strongly Agree end of the rating scale. In the second vignette, none of the specific behaviors occurred. The instructor in this case would be rated at the Strongly Disagree end of the scale. The presence or absence of each specific behavior was pointed out to the trainees after each vignette.

Part Five: Here subjects (a) observed five short videotape vignettes of different instructors lecturing, (b) answered questions on the content of the lecture and rated the instructors on their performance, and then (c) received feedback. Three different rating forms were used. The first one, used for the first vignette only, required subjects to indicate ("Yes" or "No"), that is, <u>recognize,</u> whether certain specific behaviors occurred <u>before</u> rating the related general behaviors; thus, the reporting of specific behaviors served as cues for the rating of the general behavior. The second rating form, used for the second vignette only, had fewer cues; that is, the form required subjects to <u>recall</u> the specific behaviors as reasons for their ratings of general behaviors. The third rating form was used for the remaining three vignettes. This form required subjects to rate only the three general behaviors, much like typical student "rating of instruction" forms. These rating forms are found in Appendix C.

Feedback was given immediately after subjects rated each vignette. Feedback consisted of presenting subjects with:

1. correct answers to content questions,
2. average ratings of the instructors' behavior (general and/or specific, depending upon the rating form) as determined by a group of experienced judges, and
3. the judges' reasoning behind their ratings.

Part Six: A summary which reviewed the purpose of the training program and reemphasized the general principle was given at the end of the training program. Subjects were reminded to avoid discussing the training program with other subjects in this training program and other students in the class. Also, subjects were told to observe their instructor for the next several weeks and advised that they would be rating him in one week and again in four weeks.

Descriptions of the development and pilot run of the Observation Training Program, including how the principles of training were applied and how the judges' ratings were obtained, are found in Appendix D.

Subjects in the Rater Error Training Program were trained to recognize and to avoid rating errors. Specific instructor behaviors were not mentioned. The Rater Error Training Program is similar to current training programs in performance evaluation (cf. Bernardin & Walter, 1977; Latham, Wexley, & Pursell, 1975), where the objectives are to recognize and avoid rating errors. Improvements in the training design were made, however, especially in areas of practice, feedback and transfer of training.

The Rater Error Training Program consisted of seven parts: (1) Introduction, (2) Presentation on common rating errors, (3)

Explanation of the rating scale, (4) Explanation and demonstration of the same three general instructor behaviors (but without reference to specific behaviors), (5) Practice observing and rating with feedback, (6) Review of ratings vis-a-vis rating errors discussed, and (7) Summary. The training lasted one hour.

Part One: The introduction was the same as for the Behavior Observation Training Program.

Part Two: Similar to the Behavior Observation Training Program, a short vignette was presented following the introduction in which two students were disagreeing about their ratings of an instructor on general behaviors. One student rated the instructor high on everything while the other student believed that "no one could be that good or that bad" and, therefore, always rated instructors in the middle. This scene preceded a presentation of definitions and graphic illustrations of four common rating errors: leniency, strictness, central tendency and halo. At the end of this presentation, there was a review of the four rating errors, with a short practice session in recognizing the different rating errors.

Part Three: The explanation of the rating scale was the same as in the Behavior Observation Training Program with the exception that there was no mention of specific behaviors.

Part Four: The same three general instructor behaviors were defined and illustrated using the same videotape vignettes of an actor. The only difference between the two training programs was that in the Rater Error Training Program specific behaviors associated with the general behaviors were not pointed out.

Part Five: The same five vignettes as in the Behavior Observation Training Program were also shown. The rating forms were equivalent to the third type of rating form (i.e., ratings of general instructor behaviors only) used in the Behavior Observation Training Program (see Appendix C). Feedback was given without mention of specific behaviors.

Part Six: After all five vignettes were rated and feedback was given, subjects were asked to review their ratings and to summarize them by graphing the distribution of each of the three general instructor behaviors across the five ratings. Subjects' distributions were then compared (individually) to the distributions of the judges in order for each subject to assess the likelihood he or she was committing one of the rating errors.

Part Seven: A summary of the purpose of the training program and a reemphasis of the rating errors to avoid were given at the end of this training program. As in the other training program, subjects were reminded to avoid discussing the training program with other subjects in this training program and with other students in the class. Also, they were told to observe their instructor for the next several weeks and advised that they would be rating him in one week and again in four weeks.

In summary, the two training programs were similar in terms of the training design and components--modeling, practice, feedback and transfer. The main difference was that in the Behavior Observation Training Program the objective was for subjects to base ratings of general behaviors on observations of specific behaviors (a general

principle to follow). The goal was for them to develop observation and rating skills. Further, the rating forms gradually changed, to approximate reality, from ratings of both specific behaviors and general behaviors to ratings of general behaviors only. In the Rater Error Training Program, avoidance of rating errors was emphasized by illustrating four common rating errors and giving subjects an opportunity to practice and receive feedback so they could determine the likelihood that they were making rating errors.

Subjects in the Control Group consisted of a random selection of the remaining students who came to class the day of the rating and completed the rating form along with everyone else. These students were also volunteers, but for another psychology experiment. They were not given any training or information which would otherwise make them different from typical untrained students.

Class

The independent variable Class consisted of two psychology undergraduate courses: Introductory Industrial/Organizational Psychology (Class 1) and Psychology of Advertising/Selling (Class 2). Both courses met the same three days of the week in the same room, and were large lecture classes with approximately 125 and 150 students enrolled in each, respectively. The instructor for Class 1 recently began teaching while the instructor for Class 2 has been teaching at the same university for several years. Both courses had the same prerequisites and were close with respect to median year of student (juniors).

Time Period

The effects of the treatment conditions were assessed over two time periods. Time Period 1 occurred one week (or three class sessions) after the treatment and Time Period 2 occurred four weeks after treatment.

Type of Item

There were four types of items in the rating form created by the combination of two levels of behavior, specific and general, and two levels of judgment, descriptive and evaluative. The resulting combinations, that is, the four types of items, are: specific-descriptive, specific-evaluative, general-descriptive, and general-evaluative. The content of all items was related to the instructor's ability to "maintain the attention" of the students.

A specific-descriptive item is defined as one which requires the rater (i.e., student) to report the occurrence of a specific behavior which involves little or no inference (e.g., "The instructor moved back and forth in front of the class."). A specific-evaluative item requires the rater to make a judgment about the quality or level of a specific behavior (e.g., "The instructor was above average in stating the importance of the subject matter."). A general-descriptive item is one which requires the rater to report the occurrence of a more general or more abstract instructor "behavior"; one which involves an interpretation, an inference, from the instructor's behavior (e.g., "The instructor was enthusiastic."). Finally, a general-evaluative item is one which requires the rater to make a judgment about the quality or level of a more general or more abstract instructor

"behavior" which must be inferred (e.g., "The instructor was above average in maintaining the class' attention."). A detailed explanation of the categorization of items is contained in Appendix A.

Instrumentation

Two parallel rating forms (SRI-A and SRI-B), each 24 items long, were used in the present study. SRI-A was administered at Time Period 1 and SRI-B at Time Period 2. Each rating form contained an equal number (six) of the four types of items. These items were further divided into two content areas: open communication and maintaining attention. Therefore, there were eight categories or "subtests" of three items each, although only those subtests related to 'maintaining attention' were used in the analysis of the study. For a more detailed description of the development and pilot of these rating forms, see Appendix E. The SRI-A and SRI-B rating forms are also included in this appendix.

Procedure

Subjects in the Behavior Observation Training Program met at 7:00 P.M. Wednesday or Thursday evening of the second week of the term for 1 hour 15 minutes, whereas subjects in G_2 met at 8:30 P.M. these same evenings for 1 hour. An approximately equal number of subjects were trained from both classes for each evening. Neither group knew of the differences between the contents of the two training programs. The two training programs were shown by means of videotaped presentation for standardization. A few subjects, because of scheduling difficulties, were shown their appropriate videotape separately from

their group. Subjects in both training programs were instructed to come to class regularly to observe their instructor and to rate him two times (dates were given) during the term as part of the study. There was no discussion among subjects in either group during training.

The ratings for Time Period 1 were performed at the end of the third class session (one week) after training. The entire class was administered the SRI-A rating form. Subjects in the two training programs were requested to write their student number on the rating form in order to determine their participation for the extra credit.[1] They were told that the instructor would not see the ratings. Subjects not attending class that day (there were five) were contacted immediately, that is, before the next class, to obtain their ratings. A random selection of 28 ratings in each class from the remaining students comprised the Control Group.

The ratings at Time Period 2 were performed at the beginning of the class session three weeks later (four weeks after training). The same procedure as in Time Period 1 was followed. The SRI-B form, however, was administered. Again, subjects who did not come to class at Time Period 2 (about 20) were contacted immediately to obtain their ratings. At the end of the term, all students were debriefed about the study.

Analyses

The hypotheses were tested (at $p < .05$ level of significance) by the following interaction effects from two separate four-way

---

[1] A recent study by Stone, Rabinowitz, and Spool (1977) showed that non-anonymity of student ratings of instruction does not affect ratings.

(GxCxTxI) analyses of variance (ANOVA), one for each dependent variable (interrater agreement and leniency error)--assuming a near zero correlation between the two dependent variables:

H_{1a}: Treatment Group by Type of Item (G x I) on interrater agreement

H_{1b}: Treatment Group by Type of Item (G x I) on average rating

H_{2a}: Treatment Group by Time Period (G x T) on interrater agreement

H_{2b}: Treatment Group by Time Period (G x T) on overage rating

H_{3a}: Treatment Group by Class (G x C) on interrater agreement

H_{3b}: Treatment Group by Class (G x C) on average rating

Interrater agreement was calculated by averaging the absolute values of the deviations of a subject's ratings on each item within a Type of Item from the average ratings on those same items for all subjects. Leniency error was defined as an average rating within the two most favorable ratings on the rating scale (cf. Showers, 1973). Leniency error was considered to be reduced if the average rating for training (G_1 or G_2) was significantly lower (i.e., more favorable) than the average rating for the control group (G_3).

CHAPTER IV

RESULTS

Since the two dependent variables, interrater agreement and average rating, were not significantly correlated, $r = .042$, $p < .062$, separate ANOVA's were justified. Summaries of the results of the two ANOVA's for interrater agreement and average rating are found in Table 2.

Interrater Agreement

Inspection of Table 2 reveals no significant four-way and three-way interaction effects for interrater agreement. However, three of the six two-way interaction effects, Treatment Group x Type of Item, Class x Type of Item and Time Period x Type of Item, were significant, F (6,486) = 2.283, $p < .035$, F (3,486) = 3.781, $p < .011$, and F (3,486) = 7.062, $p < .007$. Only one main effect, Class, was significant, F (1,162) = 4.078, $p < .045$.

The means and standard deviations of interrater agreement are shown in Table 3. Newman-Keuls post hoc analyses were performed on significant effects to determine where the significant effect was located (i.e., between which groups). The Newman-Keuls multiple comparison test is based on a stairstep or layer approach to significance tests (cf. Kirk, 1968). It provides a protection level lower limit of 1-$\alpha$for all ordered sets of means regardless of how many steps

Table 2

Summary ANOVA Table for Interrater Agreement and Average Rating

		Interrater Agreement		Average Rating	
Source	df	MS	F ratio	MS	F ratio
Treatment Group(G)	2	.33334	.991	.90355	.444
Class (C)	1	1.37197	4.078*	.02370	.012
G x C	2	.32967	.980	1.20675	.592
Rater (within GxC)	162	.33641	---	2.03724	---
Time Period (T)	1	.00563	.031	18.81220	20.784*
G x T	2	.17704	.971	.86377	.954
C x T	1	.44862	2.460	4.11149	4.542*
G x C x T	2	.14091	.773	.25575	.283
Rater x T	162	.18239	---	.90514	---
Type of Item (I)	3	.35353	4.077*	25.91290	95.025*
G x I	6	.19792	2.283*	.53418	1.959
C x I	3	.32782	3.781*	2.77857	10.189*
G x C x I	6	.03872	.447	.13375	.490
Rater x I	486	.08671	---	.27269	---
T x I	3	.55721	7.062*	1.95614	11.158*
G x T x I	6	.03979	.504	.19261	1.099
C x T x I	3	.13659	1.731	.04396	.230
G x C x T x I	6	.12465	1.580	.15235	.869
Rater x T x I	486	.07889	---	.17531	---

* $p < .05$

Table 3

Means[a] and Standard Deviations of Interrater Agreement

		Time Period 1				Time Period 2			
		Type of Items							
Treatment Group	Class	Specific-Descriptive	Specific-Evaluative	General-Descriptive	General-Evaluative	Specific-Descriptive	Specific-Evaluative	General-Descriptive	General-Evaluative
Behavior Observation Training	1								
	.7897 (.3759)	.7756 (.4154)	.8766 (.4499)	.7364 (.2833)	.8758 (.3317)	.7874 (.3697)	.7653 (.4204)	.7934 (.3486)	.7067 (.3743)
M = .8102	2								
(SD) = (.3676)	.8307 (.3589)	.8275 (.3875)	.7602 (.3764)	.7348 (.3204)	.8164 (.3307)	1.0425 (.3596)	.8597 (.3958)	.7992 (.2811)	.8054 (.3620)
Rater Error Training	1								
	.6996 (.3250)	.7100 (.2985)	.7134 (.3102)	.7211 (.3680)	.7569 (.4189)	.6658 (.2372)	.7874 (.3004)	.6675 (.2939)	.5748 (.3353)
M = .7625	2								
(SD) = (.3437)	.8255 (.3510)	.8350 (.2782)	.7916 (.3908)	.8478 (.3770)	.9244 (.4527)	.9039 (.3354)	.7483 (.3184)	.8291 (.3345)	.7236 (.2819)
Control	1								
	.7970 (.3818)	.8155 (.2574)	.7628 (.2949)	.7227 (.3623)	.9218 (.4826)	.7737 (.4298)	.8445 (.3665)	.6743 (.3505)	.8605 (.4452)
M = .8094	2								
(SD) = (.3693)	.8218 (.3568)	.8138 (.2772)	.7925 (.3457)	.6361 (.2426)	.9371 (.4570)	.9644 (.2962)	.7415 (.2908)	.8597 (.3400)	.8292 (.4691)
		.7670 (.3295)	.7843 (.3608)	.7267 (.3360)	.8515 (.4166)	.7423 (.3548)	.7991 (.3629)	.7117 (.3331)	.7140 (.4005)

[a]The means shown actually reflect the amount of disagreement. Therefore, the lower the mean, the greater the interrater agreement.

apart the means are. The critical value, obtained from the distribution of the studentized range statistic, for differences between means for this test varies, depending on the number of means in the set.

Newman-Keuls post hoc analyses on the Treatment Group x Type of Item effect revealed that only the Rater Error Training Group rated with significantly ($p < .01$) greater interrater agreement than the Control Group, and this difference was found only for the general-evaluative items ($M = .7449$ and .8871, respectively). None of the Treatment Groups differed from each other with respect to specific-descriptive, specific-evaluative and general-descriptive items.

Newman-Keuls post hoc analyses on the Class x Type of Item effect revealed that interrater agreement was significantly greater for Class 1 on specific-descriptive items ($M = .7547$, $p < .01$) and general-descriptive items ($M = .7828$, $p < .05$) than for Class 2 ($M = .8978$ and .8393, respectively).

As for the Time Period x Type of Item interaction effect, Newman-Keuls post hoc analyses revealed significant differences between the two time periods for specific-descriptive items and for general-evaluative items ($p < .05$ and $p < .01$, respectively). Subjects rated with lower interrater agreement at Time Period 2 than at Time Period 1 on specific-descriptive items ($M = .8563$ and .7962, respectively, and with greater interrater agreement on general-evaluative items ($M = .7500$ and .8721, respectively).

Caution should be taken in interpreting the Class and Type of Item main effects because significant interaction effects involving these independent variables exist (e.g., Class x Type of Item). That is,

the Class and Type of Item main effects are confounded and not separately estimable from the significant Class x Type of Item interaction effect. In any case, Class 1 rated with significantly ($p < .05$) greater interrater agreement than Class 2 ($M = .7621$ and .8260, respectively). As for the Type of Item main effect, Newman-Keuls post hoc analyses revealed that subjects rated general-descriptive items ($M = .7518$) with significantly greater agreement than general-evaluative items ($M = .8110$, $p < .05$) and specific-descriptive items ($M = .8263$, $p < .01$).

Average Rating

Inspection of Table 2 reveals no significant four-way and three-way interaction effects for average rating. However, three of the six two-way interaction effects, Class x Time Period, Class x Type of Item and Time Period x Type of Item, were significant, F (1,162) = 4.542, $p < .035$, F (3,486) = 10.189, $p < .0005$, and F (3,486) = 11.158, $p < .0005$, respectively. Also, two of the four main effects, Time Period and Type of Item, were significant, F (6,162) = 20.784, $p < .0005$ and F (3,486) = 95.025, $p < .0005$, respectively.

The means and standard deviations of average rating are shown in Table 4. Newman-Keuls post hoc analyses on the Class x Time Period effect revealed that Class 2 rated their instructor higher ($p < .01$) on the rating scale at Time Period 2 ($M = 3.27$) than at Time Period 1 ($M = 2.92$). As for the Class x Type of Item interaction effect, Class 1 rated their instructor significantly ($p < .01$) lower (i.e., in the more favorable direction) on specific-evaluative items ($M = 2.89$) and significantly ($p < .01$) higher (i.e., in the less favorable

Table 4

Means and Standard Deviations of Average Rating

Treatment Group	Class	Time Period 1 Type of Items: Specific-Descriptive	Specific-Evaluative	General-Descriptive	General-Evaluative	Time Period 2 Type of Items: Specific-Descriptive	Specific-Evaluative	General-Descriptive	General-Evaluative
Behavior Observation Training	1								
3.0521 (.7667)	3.0967 (.7356)	3.2262 (.6852)	2.8214 (.7561)	2.8452 (.7056)	3.3094 (.8458)	3.1547 (.7172)	2.9999 (.5367)	2.9285 (.6749)	3.4881 (.7397)
	2								
	3.0074 (.7958)	3.0357 (.7555)	2.9761 (.6965)	2.5714 (.6959)	3.0357 (.7609)	3.0953 (.8553)	3.2619 (.8575)	2.6905 (.7081)	3.3929 (.7807)
Rater Error Training	1								
3.1414 (.7527)	3.1637 (.7073)	3.2857 (.5571)	2.6786 (.5774)	2.8214 (.6509)	3.4761 (.7667)	3.2977 (.5393)	3.1428 .7393)	2.8691 (.5833)	3.7381 (.6043)
	2								
	3.1191 (.7966)	2.9524 (.6711)	2.7500 (.7678)	2.6190 (.7299)	3.2500 (.8825)	3.4287 (.6470)	3.3334 (.6542)	2.8095 (.6994)	3.8096 (.6046)
Control	1								
3.1049 (.8248)	3.0506 (.8186)	3.2143 (.7039)	2.6547 (.6758)	2.8690 (.6564)	3.2858 (1.0132)	3.2023 (.7770)	3.0238 (.8060)	2.7262 (.6354)	3.4286 (.9510)
	2								
	3.1592 (.8292)	3.0595 (.6607)	2.9525 (.7735)	2.5596 (.5522)	3.2976 (.9953)	3.2738 (.8218)	3.5952 (.6043)	2.9166 (.7460)	3.6191 (.9015)
		3.1290 (.6755)	2.8055 (.7115)	2.7143 (.6709)	3.2758 (.8793)	3.2421 (.7312)	3.2262 (.7271)	2.8234 (.6725)	3.5794 (.7792)

direction) on general-descriptive items (M = 2.84) than Class 2 (M = 3.15 and 2.69, respectively). Classes 1 and 2 did not significantly differ in average rating for specific-descriptive items (M = 3.23 and 3.14, respectively) and for general-evaluative items (M = 3.45 and 3.40, respectively).

Newman-Keuls post hoc analyses on the Time Period x Type of Item effect revealed that all items, specific-descriptive ($p < .05$), specific-evaluative ($p < .01$), general-descriptive ($p < .01$) and general-evaluative ($p < .01$), were rated significantly higher on the rating scale at Time Period 2 (M = 3.24, 3.23, 3.28 and 3.58, respectively) than at Time Period 1 (M = 3.12, 2.80, 2.71 and 3.27, respectively).

Interpretation of the Time Period main effect is questionable given its significant interaction effects with Class and Type of Item. Nevertheless, subjects in general rated their instructor significantly ($p < .0005$) higher on the rating scale at Time Period 2 (M = 3.22) than at Time Period 1 (M = 2.98). The same precaution should be taken in interpreting the Type of Item main effect because of its significant interaction effects with Class and Time Period. Newman-Keuls post hoc analyses revealed that the average rating of each type of item by all subjects differed significantly ($p < .01$) from each other. General-descriptive items were rated below the midpoint (i.e., in the more favorable, Strongly Agree direction) of the rating scale (M = 2.77), specific-evaluative items were rated at about the midpoint (M = 3.02) and specific-descriptive and general-evaluative items were rated above the midpoint, that is, in the less favorable, Strongly Disagree, direction (M = 3.19 and 3.43, respectively).

In summary, for interrater agreement, Class, Type of Item, Treatment Group x Type of Item, Class x Type of Item and Time Period x Type of Item effects were significant, and for average rating, Time Period, Type of Item, Class x Time Period, Class x Type of Item and Time Period x Type of Item effects were significant. Neither three-way interactions nor the four-way interaction were significant for both dependent variables. Therefore, only null Hypothesis 1a (Treatment Group x Type of Item effect for interrater agreement) was rejected. Separate ANOVA's were justifiable given the independence (i.e., lack of significant correlation) between the two dependent variables. Caution in interpreting significant main effects when interactions involving the independent variable are significant was suggested.

CHAPTER V

DISCUSSION

Behavior Observation vs Rater Error Training

The first research question, combined with the second and rephrased, is "Would a training program that focuses on observation skills (i.e., observing specific behaviors) increase interrater agreement and reduce leniency error more than a training program that focuses on rating errors (e.g., leniency, halo and central tendency), and if so, on which types of items?" The answer to this question is in this case "no." The major findings relevant to this question were the Treatment Group by Type of Item interaction effects for interrater agreement and average rating.

The Behavior Observation Training Program was ineffective in increasing interrater agreement across all types of items whereas the Rater Error Training Program was effective at least with respect to general-evaluative items. A possible explanation for the ineffectiveness of the Behavior Observation Training Program in increasing interrater agreement is that since subjects focused on a limited number of specific behaviors, they may have developed an observation set (i.e., were looking) for only those specific behaviors. Other relevant specific behaviors, therefore, may have been overlooked--in spite of the emphasis during training that the specific behaviors presented were only a few out of many.

A second possible explanation is that by focusing on specific behaviors, increases in variability of ratings may result if some raters observe those specific behaviors while others do not. Inspection of Table 3 reveals that interrater agreement for the Behavior Observation Training Group was, in fact, lower, although not significant, than for the Control Group across all items except general-evaluative. It has been suggested that training programs devoted to increasing trainees' responsiveness to individual differences run the risk of decreasing agreement (Crow, 1957). Findings of this sort have also been reported by Bunney and Hamburg (1963).

Interpretation of the negative results pertaining to leniency error, however, must be made cautiously. Ratings for all treatment groups, which did not differ significantly from each other, were around the midpoint of the rating scale (3.05, 3.14 and 3.10 for Behavior Observation Training, Rater Error Training and Control Group, respectively). According to the definition of leniency error (Showers, 1973), that is, average ratings within the two most favorable points on a five-point rating scale, not even the Control Group committed the leniency error. In essence, then, there was no leniency error for the Observation Training Program, and the Rater Error Training Program, to reduce.

It appears, then, that a training program, at least as conducted in the present study, which focuses on observation of specific behaviors and on ratings of general behaviors based upon those observations, does not significantly increase interrater agreement; nothing conclusive can be said about reducing leniency error. To speculate further about why the Behavior Observation Training Program was not

successful, it would be helpful to consider a model of "ideal" rater responses following three rating process steps described by Borman (1978, Note 4); (a) observing work-related behavior, (b) evaluating each of these behaviors and (c) weighting these evaluations to arrive at a single rating on a performance dimension. In relation to the first step, Borman suggested that to be accurate the rater should observe all or a perfectly representative sample of ratee behavior. The Behavior Observation Training Program only gave examples of a few representative ratee (instructor) behaviors. Presenting all or a perfectly representative sample of ratee behavior(s) may be difficult, if not impossible, with complex, general or abstract "behaviors" like enthusiasm; however, it may alleviate the potential problem described above where raters may develop an observation set for only the few behaviors presented, and, thus, possibly not observe other relevant behaviors.

The second and third steps of the rating process, according to Borman, call for agreed upon and "correct" effectiveness levels for individual behaviors and the weights assigned to those individual behaviors in developing a final picture (i.e., rating) of performance effectiveness. The Behavior Observation Training Program attempted to do this by showing videotaped examples of behaviors representative of both ends of the rating scale while at the same time pointing out the specific behaviors which occurred or did not occur as reasons why the behavior should be rated at its respective place on the rating scale. Further, feedback of the ratings by a group of expert judges on the instructors' behaviors displayed in five practice vignettes

was given to the trainees as the "correct" ratings. Trainees, therefore, had an opportunity to improve their rating accuracy.

A possible shortcoming of the Behavior Observation Training Program may be its lack of discussion among trainees. If levels and weights of behavior are to be agreed upon by trainees, then perhaps presenting trainees with the "correct" rating is not sufficient. Group discussion, where questions for clarification about the reasons for the correct ratings may be asked and answered, may serve to clarify personal biases and help to establish group norms regarding ratings. The result, hopefully, would be an increase in interrater agreement and a decrease in rating errors. But group discussion of ratings is still not enough, as Bernardin (1978), Borman (1975) and Latham, Wexley, and Pursell (1975) found. A workshop, like the Behavior Observation Training Program, where trainees have the opportunity to practice and receive feedback, among other things, must be conducted to obtain significant results. However, when evaluating training programs involving group discussion, the researcher must use the appropriate unit of analysis--the group. One way to achieve reasonable statistical power while meeting the need to have group discussion is to divide the training group into small, independent discussion groups. These small discussion groups then become the unit of analysis--rather than the total training group ($\underline{n}$ = 1). The researcher should be aware of the possibility of different group norms developing for each of the small groups.

Rater Error Training

The present study found the Rater Error Training Program to be effective in increasing interrater agreement, but only with general-evaluative items. This finding may explain the generally negative findings found with other rater error training programs attempting to increase interrater agreement (e.g., Bernardin & Walter, 1977; Borman, 1975; Borman, Note 3). These studies evaluated their training programs without considering the interactive effect of types of items. As in the present study, training, when evaluated across all types of items combined, produces non-significant results. However, if types of items were taken into consideration, especially where general-evaluative items are used in the rating form, significant results may be obtained. Noteworthy is the fact that in both the Bernardin and Walter (1977) and the Borman (1975) studies, behaviorally anchored rating scales/forms were used. This type of rating form consists entirely of specific-descriptive items--where training (in the present study as well as in those studies) was shown to have no effect.

Other Results

Other results obtained in the dissertation are worth discussing. The non-significant Treatment Group by Time Period interaction for interrater agreement suggests that the interrater agreement on general-evaluative items among subjects in the Rater Error Training held up over time. However, the time period was only one month and interrater agreement at some later time is not known. Recent findings have suggested that observer agreement decreases over time and has been referred to as the "observer drift" phenomenon (Johnson & Bolstad,

1973; O'Leary & Kent, 1973; Reid, 1970; Reid & DeMaster, 1972; Romanczyk, Kent, Diament, & O'Leary, 1973; Talpin & Reid, 1973; Reid, Note 5).

The significant Time Period by Type of Item interaction effect on interrater agreement may suggest an alternative explanation (other than due to training) to the increase in interrater agreement for general-evaluative items. Specifically, at Time Period 2 general-evaluative items were rated with greater interrater agreement than at Time Period 1, while interrater agreement decreased for the other types of items. The nature of general-evaluative items such that they require an integration of a whole set of observations, which with the addition of more observations agreement among raters would increase. The increase found in interrater agreement for general-evaluative items, therefore, may be more of a function of time rather than training. This alternative explanation appears consistent with the Spearman-Brown correction formula in reliability which, stated simply, says that more items (observations) increase a test's (rating's reliability.

The Treatment Group by Class Interaction effect was also not significant for interrater agreement. This means that the Treatment Groups did not differ in interrater agreement with respect to the class from which subjects came. In other words, results of Rater Error Training may be considered consistent across classes; at least the two classes involved in this study. On the other hand, these courses consisted of students from many majors (e.g., psychology, communications, business, human ecology, etc.), ranging from freshmen

to seniors. Therefore, it may be possible to generalize across some groups of undergraduate students. Further, since instructors are confounded with classes, it may be that training effects are generalizable across instructors as well.

Regarding generalizability, the present study dealt with student ratings of instruction. This situation involves a group of raters (students) who observe one ratee (instructor) for short periods over time. In addition, the main concern of the students is to learn the content of the lecture; observation of instructor behaviors has minimal priority. To generalize beyond this setting, therefore, may be questionable. For example, supervisors who rate more than one subordinate observe behaviors during less intense observation periods. Further, supervisors' main concern is with the behavior of their subordinates. Thus, not only is this observation setting different, but so is the motivation to observe. Generalization should be limited to settings where observation of behavior is done in a controlled setting and is considered important, such as in assessment centers.

Practical Significance of Training

There is one other general finding which has significant implications. When the total picture of results is looked at, it becomes apparent that training of raters was effective for only a small part of the overall ratings. None of the training program groups were superior to the control group in reducing rating errors. Moreover, only the Rater Error Training Program was effective in increasing interrater agreement, and this was only for general-evaluative items. Pilot data suggested that general-evaluative items initially have the

greatest chance for increase in interrater agreement. But, the percent of variance in interrater agreement accounted for by the Treatment Group x Type of Item interaction is extremely small (eta = 2.6%).

These results at first seem in contrast to the general positive findings claimed by studies evaluating training programs for raters in performance evaluation. However, closer examination of the results of these other studies does not suggest such an overall positive effect of training, at least from a practical significance standpoint. For example, interrater reliability in Borman's (1975) study actually decreased significantly after a short training (i.e., presentation) session. In the studies by Bernardin and Walter (1977) and by Borman (Note 4), the training programs investigated were not significantly different from the control group in interrater agreement. Also, in Bernardin and Walter's study, there was no practical difference between a comprehensive training program (average rating was 4.4) and a control group (average rating was 4.8, with an average standard deviation of 1.06) in reducing leniency error. In another study, Bernardin (1978) found similar results when comparing the same training program with a control group ($\underline{M}$ = 4.3 and 4.8, respectively, and with an average standard deviation of 1.33). The average rating of the control groups (4.8 on a seven-point rating scale) in these latter two studies also suggests that leniency error was not a problem to begin with). Therefore, results of other studies in conjunction with results of the present study question the practical significance of training in increasing interrater agreement (at least for items other than general-evaluative) and in reducing leniency error.

Summary and Conclusion

In summary, training directed at recognizing rating errors only was shown to be effective in increasing interrater agreement when rating forms contain general-evaluative items. The effects of training were also consistent over time and across classes/instructors. Neither training program, however, differed from the control group in average rating--indicating no effect in reducing leniency error.

What do these results mean, then, with respect to the training of raters to evaluate performance? It seems that training raters to recognize rating errors is only helpful/necessary if the rating form contains general-evaluative items and interrater agreement is of concern. Directing training toward observing specific behaviors, however, is probably not effective in increasing interrater agreement and in reducing leniency error, and may have the potential of actually decreasing interrater agreement. If rating forms contain items other than general-evaluative, the practical benefit of conducting training would be questionable. Furthermore, the existence of leniency error as a problem in performance ratings should be determined beforehand.

Future Research

Research in the future should focus on three areas: (2) improvement of behavior observation training, (b) determination of the practical significance of rater training, and (c) the measurement of item types. With regard to the first area, research on behavior observation training should still be pursued. Improvements to the present behavior observation training program were suggested. Briefly, they were (a) include all or a perfectly representative sample of behavior in the

training and (b) hold small group discussions about ratings, with the unit of analysis being the small groups.

As for the second area, Bernardin (1978) stated that before the practical significance of training can be addressed, the validities of the ratings must be estimated. Field tests, he states, would be impossible. However, laboratory tests may prove to be a useful starting place. For example, patterned after the Borman (1978) study, ideal conditions could be set up for trainees to rate instructors to determine the highest level of interrater agreement achievable for the different types of items. In this way, if negative results occur, one could determine whether or not subjects were already at their maximum--thus training could not improve interrater agreement to any greater extent--or that training was not effective. An increase in interrater agreement, by the way, may not be practical if it would not change the ratee's acceptance of his or her performance ratings. A study to determine the relationship between interrater agreement and credence in the rating appears warranted since high interrater agreement is assumed desirable because of assumed higher acceptance of ratings (in addition to higher validity). It should be noted, however, that 100 percent agreement is neither desirable, because of the "attenuation paradox," nor expected, because of many intervening variables affecting interrater agreement on performance ratings. Comparisons between training programs, as in the present study and as in Bernardin (1978) and Latham, Wexley, and Pursell (1975), should be continued to determine the practical significance training programs, in improving ratings of performance, varying not only content but length. Recent findings

(Bernardin, 1978), for example, have questioned the practical improvement of ratings by a comprehensive training program one hour in length over a five minute "awareness" session.

Future research should also focus on the categorization/measurement of different types of items as a means of defining rating task difficulty. It was shown in the present study that Type of Item, as an independent variable, significantly interacts with other independent variables (e.g., Treatment Group, Class and Time Period). The present study used a subjective categorization of "types" of items: judges' agreement according to a definition which categorized an item into one of four types. Much information is lost using non-continuous measurement (e.g., categorization). An index measuring, on a continuous scale, degree of behavior-specificity and degree of judgment might be preferable. However, it should first be demonstrated whether or not improvement in the calibration of item types is possible and practical.

In conclusion, if research is to continue in the area of training of raters to evaluate performance, the above three areas of research must be addressed. Researchers, however, should be forewarned of probably drawing conclusions similar to findings in research on behavioral anchored rating scales--little practical improvement in the quality of ratings over what already exists (Borman & Dunnette, 1975). The questions to be answered next are, "Could the problem with improving rating scales be occurring with the training of raters?" That is, "Is training raters to increase interrater agreement and reduce rating errors beyond a short (e.g., five minute) 'awareness' session practical, worth the investment?"

REFERENCE NOTES

REFERENCE NOTES

1. Spool, M.D. Effects of inference, judgment and content on degree of agreement in student ratings of instruction. Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada, 1978.

2. Burnaska, R. F. Personal Communication, June 14, 1977.

3. Bernardin, H. J. Personal Communication, April 19, 1977.

4. Borman, W. C., & Rosse, R. L. Format and training effects on rating accuracy and rater errors. Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada, 1978.

5. Reid, J. B. The relationship between complexity of observer protocols and observer agreement. Paper presented at the annual meeting of the American Psychological Association, Montreal, Canada, 1973.

REFERENCES

REFERENCES

Becker, W. C. The relationship of factors in parental ratings of self and each other to the behavior of kindergarten children as rated by mothers, fathers, and teachers. Journal of Consulting Psychology, 1960, 24, 507-527.

Bergquist, W. H., & Phillips, S. R. Components of an effective faculty development program. Journal of Higher Education, 1975, 46, 177-211.

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

Boehm, A. E., & Weinberg, R. A. The classroom observer: A guide for developing observation skills. New York: Teachers College Press, 1977.

Borg, W. R., & Gall, M. D. Educational research: An introduction. New York: David McKay Company, Inc., 1971.

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. Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 1978, 63, 135-144.

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

Burnaska, R. F. The effects of behavior modeling training upon managers' behaviors and employees' perceptions. Personnel Psychology, 1976, 29, 329-335.

Byrne, D. Assessing personality variables and their alteration. In P. Worchel and D. Byrne (Eds.), Personality change. New York: Wiley, 1964, 38-68.

Cartwright, C. A., & Cartwright, G. P. Developing observation skills. New York: McGraw-Hill Book Company, 1974.

Cohen, J., & Humphreys, L. G. Memorandum to faculty. University of Illinois, Department of Psychology, 1960. Reported in F. Costin, W. T. Greenough, and R. J. Menges, Student ratings of college teaching: Reliability, validity and usefulness. Review of Educational Research, 1971, 41, 511-535.

Conrad, H. S. The personal equation in ratings. I. An experimental determination. Pedagogical Seminary and Journal of Genetic Psychology, 1932, 41, 267-292.

Costello, A. J. The reliability of direct observations. Bulletin of the British Psychological Society, 1973, 26, 105-108.

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity and usefulness. Review of Educational Research, 1971, 41, 511-535.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generaliability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Davis, R. H., Alexander, L. T., & Yelon, S. L. Learning system design: An approach to the improvement of instruction. New York: McGraw-Hill Book Company, 1974.

Doyle, K. O., Jr. Student evaluation of instruction. Lexington, Mass.: Lexington Books, 1975.

Dunnette, M. D. Personnel selection and placement. Belmont, Calif.: Wadsworth, 1966.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Engelhart, M. D. A method of estimating the reliability of ratings compared with certain methods of estimating the reliability of tests. Educational and Psychological Measurement, 1959, 19, 579-588.

Finn, R. H. A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 1970, 30, 71-76.

Finn, R. H. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. Educational and Psychological Measurement, 1972, 32, 255-265.

Gellert, E. Systematic observation: A method in child study. Harvard Educational Review, 1955, 25, 179-195.

Goldstein, A. P., & Sorcher, M. Changing supervisor behavior. New York: Pergamon Press, 1974.

Guilford, J. P., & Jorgensen, A. P. Some constant errors in ratings. Journal of Experimental Psychology, 1938, 22, 43-57.

Heilman, J. D., & Armentrout, W. D. The rating of college teachers on ten traits by their students. Journal of Educational Psychology, 1936, 27, 197-216.

Heyns, R. W., & Lippitt, R. Systematic observational techniques. In G. Lindzey (Ed.), Handbook of social psychology (Vol. 1). Cambridge, Mass.: Addison-Wesley, 1954, 370-404.

Hinrichs, J. R. Personnel training. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976, 829-860.

Holdaway, E. A. Different response categories and questionnaire patterns. Journal of Experimental Education, 1971, 40, 57-60.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: methodology, concepts, and practice. Champaign, Ill.: Research Press, 1973.

Jones, R. R., Reid, J. B., & Patterson, G. R. Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 3). San Francisco: Jossey-Bass, 1975, 42-95.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. Student evaluations of teaching: The generalizability of class means. Journal of educational measurement, 1976, 13, 171, 183.

Kaplan, A. The conduct of inquiry. San Francisco: Chandler, 1964.

Kirkpatrick, D. L. Techniques for evaluating training programs. Training and Development Journal, 1959, 13, 3-9, 21-26.

Kulik, J. A., & McKeachie, W. J. The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), Review of research in education (Vol. 3). Itasca, Illinois: F. E. Peacock Publishers, Inc., 1975, 210-240.

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.

Levine, J., & Butler, J. Lecture versus group discussion in changing behavior. Journal of Applied Psychology, 1952, 36, 29-33.

McCormick, E. J., & Tiffin, J. Industrial psychology (6th edition). Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1974.

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

O'Leary, K. D., & Kent, R. Behavior modification for social action: Research tactics and problems. In L. A. Hamerlynck, L. C. Handy, & H. E. Mash, (Eds.), Behavior change: Methodology concepts and practice. Champaign, Illinois: Research Press, 1973.

Reid, J. B. Reliability assessment of observation data: A possible methodological problem. Child Development, 1970, 41, 1143-1150.

Reid, J. B., & DeMaster, B. The efficacy of the spot-check procedure in maintaining the reliability of data collected by observers in quasi-natural settings: Two pilot studies. Oregon Research Institute Research Bulletin, 1972, 12, No. 8.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. Measuring the reliability of observational data: A reactive process. Journal of Applied Behavior Analysis, 1973, 6, 175-184.

Rosenshine, B. Teaching behaviors and student achievement. Atlantic Highlands, N. J.: Humanities Press, Inc., 1971.

Rosenshine, B., & Furst, N. Research on teacher performance criteria. In B. O. Smith (Ed.), Research in teacher education. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1971, 37-72.

Showers, B. H. Alternative response definitions in instructional rating scales. Unpublished doctoral dissertation, Michigan State University, 1973.

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.

Spool, M. D. Training programs for observers of behavior: A review. Personnel Psychology, in press.

Stewart, C. T., & Malpass, L. F. Estimates of achievement and ratings of instructors. Journal of Educational Research, 1966, 59, 347-350.

Stone, E. F., Rabinowitz, S., & Spool, M. D. Effect of anonymity on student evaluations of faculty performance. Journal of Educational Psychology, 1977, 69, 274-280.

Talpin, P. S., & Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. Child Development, 1973, 44, 547-554.

Walter, H. M., & Gilmore, S. K. Placebo versus learning effects in parent training procedures designed to alter the behaviors of aggressive boys. Behavior Therapy, 1973, 4, 361-377.

Weick, K. E. Systematic observational methods. In G. Lindzey and E. Aronson (Eds.), The handbook of social psychology. Reading, Mass.: Addison-Wesley Publishing Co., 1968 (Vol. 2), 357-451.

Weinrott, M. R. Observation training and practice: Effects on perception of behavior change. Unpublished doctoral dissertation, McGill University, 1975.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. Training interviewers to eliminate contrast effects in employment interviews. Journal of Applied Psychology, 1973, 50, 233-236.

APPENDICES

APPENDIX A

CATEGORIZATION OF ITEMS

APPENDIX A

Categorization of Items

A pool of 276 items was formed from which items were categorized into one of the four types. Most of the items in the pool were selected from existing student rating of instruction forms with proven (via research studies) validity and reliability (e.g., Cornell Diagnostic Observation and Reporting System and University of Illinois' CEQ). Some of the items were rewritten for one or more of the following reasons: (1) to allow for a general format (e.g., beginning an item with an action verb and having "the instructor" the subject), (2) to increase the number of possible evaluative items to be categorized by adding or making more clear a level of quality (e.g., "very well" or "above average") to the instructor's behavior, (3) to permit the formation of two parallel forms (A and B) by writing similar items within the same category, (4) to make the items gramatically correct and easier for undergraduates to read and understand, and (5) to assure that if an inference were to be made when responding to an item, the inference was about the instructor's behavior and not about the content (e.g., whether or not the subject matter was interesting). Items which measured student outcomes or which were highly dependent on individual differences (e.g., "Made difficult topics easy to understand") were not included. Added to the original pool of 276 items were 30 repeat items

randomly chosen for a reliability check. This made the total number of items to be categorized 306.

Six graduate students from Learning and Evaluation Service Department, each having research or "hands-on" experience with student ratings of instruction, served as judges to categorize the pool of items into one of four types. These types of items, formed by the combination of two levels of behavior (specific and general) and two levels of judgment (descriptive and evaluative), are: (1) specific-descriptive (S-D), (2) specific-evaluative (S-E), (3) general-descriptive (G-D) and (4) general-evaluative (G-E). An S-D item was defined as one which required the observer (i.e., student) to report the occurrence of a specific behavior which involves little or no inference (e.g., "The instructor moved back and forth in front of the class."). An S-E item required the observer to make a judgment about the quality or level of a specific behavior (e.g., "The instructor was above average in stating the importance of the subject matter.") A G-D item is one which required the observer to report the occurrence of a more general or more abstract instructor "behavior"; one which involves in interpretation, an inference, from the instructor's behavior (e.g., "The instructor was enthusiastic."). Finally, a G-E item is one which required the observer to make a judgment about the quality or level of a more general or more abstract instructor "behavior" which must be inferred (e.g., "The instructor was above average in maintaining the class' attention.").

Instructions with definitions of each category (at end of appendix) were given to the judges. Answer sheets with the four

possible categories (S-D, S-E, G-D, and G-E) listed next to the item number were also provided. The judges categorized the items independently of each other. The average length of time to do this task was about one hour.

The judges' responses were then tallied. There were eighty-nine items with 100 percent (six out of six) agreement. The number of items falling under each category was determined. The next level of agreement (five out of six, or 83 percent) had the following stipulation: the disagreement among judges had to be in only one dimension--either level of behavior or level of judgment, but not both. For example, items which were categorized as S-D and G-E or G-D and S-E were eliminated. Seventy-four items were categorized with 83% agreement. The lowest level of agreement considered, with the same criterion as the previous level (i.e., categorization across no more than one dimension), was 67 percent (four out of six). There were 46 of these items.

Table A1 shows the distribution of items across types and amount of agreement. In total, 209 out of 276 items, or about 76%, were able to be categorized. Of the 30 repeat items, 25 (or 83%) were categorized the same as the first time.

Table A1

Categorized Items

		Type of Item				
		S-D	S-E	G-D	G-E	n
Level of Agreement	100% (6/6)	49	7	17	16	89
	83% (5/6)	32	4	21	17	74
	67% (4/6)	13	11	14	8	46
	n	94	22	52	41	209

Instructions to Judges

Attached is a list of statements/items commonly found in a student rating of instruction form. Your task is to put these items into one of four categories: specific-descriptive; specific-evaluative; general-descriptive; and general-evaluative.

A specific-descriptive (S-D) item is one which requires the observer (the student filling out the instruction rating form in this case) to report the occurrence of a specific behavior which involves little or no inference. Examples of specific-descriptive items are: "The instructor paced up and down in front of the class as he spoke" or "The instructor told students before he defined a concept that he was going to give a definition."

A specific-evaluative (S-E) item is one which requires the observer to make a judgment about the quality or level of a specific behavior. For example, a specific-evaluative item is one which states how well the instructor did something like "The instructor was very effective in questioning students about the subject matter."

A general-descriptive (G-D) item is one which requires the observer to report the occurrence of a more general or more abstract instructor "behavior," one which involves an interpretation, an inference, from the instructor's behavior. An example would be "The instructor was self-confident." One cannot see an instructor's self-confidence; one has to infer it. In other words, self-confidence is not a behavior but rather an inference made from a number of behaviors.

A general-evaluative (G-E) item is one which requires the observer to make a judgment about the quality or level of a more general

or more abstract instructor "behavior" which must be inferred. An example of a general-evaluative item is one which asks students to judge the quality of the instructor's relationship with students (e.g., "The instructor did a good job of really relating to you").

Below are examples of the same item worded in each of the four (S-D, S-E, G-D and G-E) ways.

(S-D)	S-E	G-D	G-E	Stated his personal experiences.	SA	A	N	D	SD
(S-D)	S-E	G-D	G-E	Showed examples of his hobbies.	SA	A	N	D	SD
S-D	(S-E)	G-D	G-E	Was exceptionally good in stating his personal experiences.	SA	A	N	D	SD
S-D	S-E	(G-D)	G-E	Revealed his personal experiences.	SA	A	N	D	SD
S-D	S-E	G-D	(G-E)	Did a good job of revealing his personal experiences.	SA	A	N	D	SD

Use the answer sheet provided and circle the appropriate category letters for each item (S-D = specific-descriptive; S-E = specific-evaluative; G-D = general-descriptive; G-E = general-evaluative). Make sure all items are clearly marked. Read each item carefully because some items which may appear similar are not necessarily the same. Also, do not look back at previous answers.

Should you have any questions, ask me or call me at 3-4645. Your cooperation is greatly appreciated.

APPENDIX B

SCRIPTS FOR TRAINING PROGRAMS

APPENDIX B

TRAINING MANUAL

Observation and Rater Error Training Programs

Introduction

This manual provides guidance for trainers training students to rate general instructor behaviors. The goal of training is to enable the participants (students) to increase agreement with each other and reduce rating errors on ratings of items about general instructor behaviors.

Trainers should follow as closely as possible the procedure and statements presented in this manual to assure standardization. Fine variations of the training material is acceptable when attempting to present the material in a comfortable manner.

Important Considerations

1. It is important to keep the training directed toward the individual. To have the unit of analysis at the individual level, the trainees must not interact during training. This, in particular, concerns the feedback part of training. After the trainees practice observing and rating the instructor in the vignettes, the trainer should announce the judges' answers and reasons for those answers. Neither the trainees' responses nor the correct answers should be discussed. Questions about

procedures, examples, etc., can be answered by the trainer in front of the group, but no trainee-trainee interaction should occur during the practice-feedback portion of training.

2. Throughout the training, the trainer should periodically make reference to or remind the trainees of the general principle: (1) base ratings of general behaviors on observations of related specific behaviors or (2) avoid rating errors.

Training Schedule

This workshop is designed for one session, one hour long. Training is, therefore, conducted to a pre-specified time limit. When trainees enter, they should sign their name and student number on a sheet of paper and be given a handout explaining the purpose of training and giving an overview. When trainees leave, they should be given a 'reminder' handout-sheet.

Instructional Conditions

For best results, the following conditions should be met when conducting this workshop:

1. Allow for more time than required; provide leeway for set-up and clean-up.
2. Have sufficient (i.e., more than necessary) number of workshop materials.
3. Use a classroom with privacy, chalkboard, separate chairs and writing extensions, and a screen for an overhead projector.
4. Use a videotape system to present training program. The videotape system must be set up to turn "on" and "off" at various times during training (e.g., practice rating).

Instructional Material

1. This training manual handout
2. Handout explaining purpose of training, ground rules and overview
3. Rating forms-packet
4. Videotape cassette with playback machines and television
5. Handout reminding students to avoid discussing training program with others, to attend class regularly and to rate the instructor in one week and again in four weeks.

Objective

To increase interrater agreement and reduce rating errors when students rate general instructor behaviors.

Improving Ratings of Instruction

The purpose of today's session is to improve student ratings of instruction, like the ones you fill out at the end of the term. Most of the session is on videotape and will last about one hour.

After an introduction, you will learn a method of rating instructors. This method emphasizes the consideration of specific behaviors of an instructor rather than just your general impressions. You will watch on videotape an instructor demonstrating three general instructor behaviors characteristic of instructors:

1. Making the subject matter relevant to students
2. Helping students keep their attention on the subject matter
3. Being enthusiastic when lecturing

You will also get a chance to practice rating five different instructors on these general behaviors. The purpose of these practice ratings is to help improve your ratings, not to evaluate you. You will get some feedback on how a group of experienced observers rated the instructors.

To maintain independence among participants, it will be necessary to avoid discussion during today's session, particularly during the practice ratings and feedback.

Next week on Wednesday, April 19, at the end of class you will be asked to rate your instructor. You will also rate your instructor in about a month, on Friday, May 12.

At the end of today's session I will give you a handout with these dates as a reminder.

Your cooperation is greatly appreciated.

SCRIPT FOR BEHAVIOR OBSERVATION TRAINING PROGRAM

Introduction

Attention Getter

"Have you ever been asked to rate your instructor on something general like 'organization' and then ask yourself, 'I wonder what they mean by that?' [pause] Then, you find out that someone else rated the instructor high while you rated him low."

"It is not uncommon for students to disagree with each other when rating instructors, and as a result, instructors have a hard time interpreting their ratings. Which students should they believe?"

Purpose

"It has been said that the quality of student ratings of instruction depend upon the ability of students to accurately observe and rate their instructor. If students can be trained to observe their instructor, then the quality of their ratings should improve."

"We are investigating whether or not this is true. What you will learn from this training session should really help you rate your instructors. This, then, should really have an impact on how instructors look at the ratings they get."

"Today I will train you in how to accurately observe and rate three (3) general instructor behaviors."

"After today's session, I'll want all of you to pay particular attention to similar teaching behaviors on the part of your instructor back in your classroom. Next week and again in four weeks you will rate your instructor."

"I'll tell you more about this later, but you should be thinking about it while you go through the training."

Ground Rules

"We're going to be running this training during this week with other students in your class. Therefore, we ask you not to discuss tonight's session with other students in your class because that might influence or bias them."

"Similarly, because each of you will be rating your instructor one week and four weeks from now, please do not discuss tonight's session with others within this group during and after tonight's session."

"To summarize, we request that you do the following things:

1. Learn what is taught today as best you can.
2. Avoid discussing the content of today's session with any students in this group and other students in the class.
3. Observe your instructor for the next several weeks--this means you should attend class regularly.
4. Rate your instructor in class Wednesday, April 19.
5. Rate your instructor again in class on Friday, May 12.

Overview

"Today's session will last about an hour. I will first review with you an approach on how to rate general instructor behaviors, that is, base your ratings of general behaviors on observations of specific behaviors. Then we will go over three (3) general instructor behaviors that are important:

No. 1. making the subject matter relevant

No. 2. helping you keep your attention on the subject matter

No. 3. being enthusiastic when lecturing.

To better understand each of these general behaviors, you will see short examples of the specific behaviors that are related to them. Then you will watch a series of short scenes of different instructors and you will practice rating these instructors on the 3 general behaviors. Finally, I will give you feedback after each practice rating on how a group of experienced observers rated that instructor.

Rating General Behaviors

"Did you ever get into a discussion with other students about your instructor and found that they had quite a different view?"

"Consider, for example, the following situation:

A: Boy, this instructor isn't very organized.

B: I don't know. He sure looks it to me. Didn't you notice when he told us what will happen in the next couple of classes; and what about those handouts?"

A: I guess I didn't think of those things. But still, he didn't tell us where we are in relation to the rest of the course. I still think he's unorganized."

"Why do you think there can be such a wide difference in their views? Is one right and the other wrong, or can both be right?"

"Statements like "The instructor was organized" require students to respond to a general behavior of an instructor."

"However, organization, for example, is not behavior. We can only infer that it exists from observing specific behaviors, like showing an outline and presenting material in sequence. Specific behaviors can be seen, observed, by everyone. So, a general behavior like 'organization' is, by itself, not a behavior but is made up of specific behaviors."

"A lot of times students react differently to the same instructor. Granted, there may be real differences seen by students, but often the differences found are because of different perspectives or interpretations of the general behavior the students are asked to respond to."

"These different perspectives may exist because of different backgrounds or experiences. As a child, for example, you might have been told that a certain individual was exciting--but you were not told why. Later you might have heard that another person was a bore. [pause] It was up to you to figure out why. [pause] More than likely you tried to recall what the person did and then attribute those behaviors to the general description of 'exciting' or 'boring.' [pause] Because it is difficult to recall every behavior, we select only a few. If people recalled or selected different behaviors, then their general reactions to the same person will be different. The question still remains, 'Is the person exciting or not?'"

"To overcome this problem, it has been strongly recommended that ratings on very general behaviors, such as 'exciting' or 'organized' be based on observations of specific behaviors. Further, if students recognize the same specific behaviors as they relate to a general behavior, then there should be greater agreement among them. The more agreement among students, the greater the likelihood the instructor will use the ratings to improve the lecture."

"So, I would like you to use the following procedure when rating instructors on general behaviors:

> First, observe those specific behaviors which are related to the general behaviors you are to rate.

In the next part, I will tell you some of the specific behaviors which relate to each of the general behaviors.

> Second, try to recognize situations when these specific behaviors could occur; that is when opportunities for their use arise.
>
> Third, consider the number of occurrences and/or non-occurrence of the specific behaviors; then rate the instructor on the general behavior (i.e., the extent to which you agree that the general behavior is characteristic of the instructor."

"To summarize, ratings of general behaviors should be based on what the instructor actually did or did not do; not what you thought he did or what you expected him to do. We'd like you to base your ratings of general behaviors on observations of the presence or absence of relevant [pause] specific [pause] behaviors. Such a method of observation should obtain greater agreement among the students."

Explanation of the Rating Scale

"The rating format for the 3 general behaviors you will be rating will look like this:

The instructor:

1. Made the subject matter relevant. SA A N D SD
2. Helped you keep your attention on the subject matter. SA A N D SD
3. Was enthusiastic when lecturing. SA A N D SD

The letters for the rating scale stand for the following:

SA = if you strongly agree with the statement

A = if you agree with the statement

N = if you neither agree nor disagree

D = if you disagree with the statement

SD = if you strongly disagree with the statement

"When you read the statement about the instructor, you are to respond by circling the letter representing the extent [pause] to which you agree with it."

"Since we believe that ratings of general behaviors should be based on specific behaviors, if the instructor does many specific behaviors related to the general behavior, then you should circle SA or A, depending upon the number of specific behaviors occurring. However, if the instructor does only a few or none of the relevant specific behaviors, then you should rate toward the D or SD end of the scale, again depending upon the number of specific behaviors occurring or not occurring. It is up to you to decide how many relevant specific behaviors are necessary to agree with the statement."

"To help you get a better understanding of all this, I am going to give you examples of some specific behaviors related to each of the 3 general behaviors you will be rating the instructors on. It is important to understand that there can be many specific behaviors related to one general behavior, but we will only consider two or three. Also, these general behaviors are not totally separate. For example, being enthusiastic may help you keep your attention on the subject matter. However, those two general behaviors are not the same--each has specific behaviors different than the other."

Explanation and Demonstration of the Specific Behaviors Related to Three General Instructor Behaviors

"For each general behavior, I will be showing you videotapes of an instructor, first where all the specific behaviors are present and second where none of the specific behaviors are present. In the first case, you would rate the instructor at the Strongly Agree & Agree end of the rating scale. In the second case, you would rate the instructor at the Strongly Disagree & Disagree end of the scale. For each general behavior, see if you can spot the difference between the instructor at the SA end of the scale and at the SD end of the scale."

"For the first general behavior, 'The instructor made the subject matter relevant,' two related specific behaviors are:

1. Stated why the material is being presented (e.g., the importance of the topic).
2. Stated how the content relates to your interests, background or activities.

"When a topic or subject matter is being presented, it is important for students to know why it is being presented; that is, why the topic is important."

"It has also been found that students can learn better if the subject matter is made relevant or meaningful to them. In this case, even if the material seems generally interesting, students must be directly told [PAUSE] how the content relates to them [PAUSE] personally. [PAUSE]

"I will now show you two videotaped lectures by the same instructor on the topic of 'motivation in organizations.' The instructor is lecturing to a group of undergraduate students in management. In the first videotape, both specific behaviors are present. In this case, the instructor would be rated toward the SA and A end of the rating scale. See if you can recognize the specific behaviors.

[SHOW 1st VTR: EXAMPLE]

"The instructor you just watched made the subject matter relevant. Let's look at the parts of the lecture where the specific behaviors were demonstrated."

"Remember when the instructor stated the importance of the topic?"

[SHOW S.B. 1 SEGMENT]

"What about the instructor directly stating how the content related to the students? What did he say?"

[SHOW S.B. 2 SEGMENT]

"Now, let's watch the same instructor but this time he doesn't do the specific behaviors. See if you can determine the difference between this scene and the first one in terms of making the subject matter relevant."

[SHOW 1st VTR: NON-EXAMPLE]

"This time the instructor didn't state the importance of the topic and he didn't state how the subject matter was related to the students. Even though the topic may be generally interesting to everyone, the instructor did not directly state how the topic was related to the students, individually and personally. The instructor in this case should be rated toward the SD and D end of the scale."

"For the second general instructor behavior, 'Helped you keep your attention on the subject matter,' we will consider 3 specific behaviors. Remember, there may be other specific behaviors, but we will only consider 3 of them. They are:

1. Made a statement to grab students' attention (e.g., a puzzling question, a contradictory or powerful statement, etc.).
2. Showed where each point fits into an outline, especially as he comes to it.
3. Stated at least one meaningful example or illustration for each major point.

"In a presentation on any topic, the instructor can do several things in the beginning to get the attention of the students. The instructor can ask a puzzling question, make a seemingly contradictory statement or even a powerful statement."

"To maintain students' attention throughout a presentation, an instructor should present some kind of organization or outline, as well as follow and refer to it periodically."

"When complex material is being presented, a student can easily get lost and/or lose attention unless meaningful (i.e., relevant) examples or illustrations are made for each major point."

"In the next two videotapes you will see the same instructor first demonstrate a lecture with all the specific behaviors present (and therefore would be rated toward the SA and A end of the scale) and next demonstrate a lecture with none of the specific behaviors present (and would be rated toward the SD and D end of the scale)."

"In the first videotape coming up, see if you can identify the specific behaviors."

[SHOW 2nd G.B.: EXAMPLE]

"The instructor you just watched helped students keep their attention on the subject metter. Let's consider the specific behaviors he demonstrated."

"Remember <u>what</u> the instructor said near the beginning which got the students' attention? He described two workers and then he said: [QUOTE] "Why do you think there was such a wide difference between these two workers? Was the first motivated more? Or, was it that the 2nd worker did not have the ability to do the work?"

"What about the instructor referring to an outline as he came to a point? Remember that?"

[ACT OUT] He said, "Today, as you can see from our outline we will be discussing the effect of system of pay on the motivation of workers."

"Do you remember the example he gave? He told about the reason the first worker worked so hard."

"Now let's watch the same instructor but this time he doesn't do the specific behaviors. He would be rated toward the SD and D end of the scale. See if you can determine the difference between this scene and the previous one in terms of helping the students keep their attention on the subject matter."

[SHOW 2nd VTR: NON-EXAMPLE]

"This time the instructor didn't make a statement to grab the students' attention. Neither did he show where each point fits into an outline nor did he give an example or illustration of the major points being made. The instructor in this case should be rated toward the SD and D end of the scale."

"For the third general behavior, 'Was enthusiastic when lecturing,' we will consider only 3 out of several specific behaviors. These specific behaviors are:

1. Varied voice (volume, speed, pitch).
2. Varied movement/activity; did not just remain still.
3. Presented subject matter with personal examples from his/her own experiences.

"Whenever presenting, an instructor who changes his voice pattern (like varying loudness, speed or tone) will captivate students' interests. Such a presentation appears to be stimulating and enthusiastic."

"Also whenever presenting, the instructor can move around (that is, the position standing in, arms or hands, or any combination of body movement) or use different instructional devices, such as charts. So long as such movement is not distracting, movement or different activities make a presentation or lecture appear more enthusiastic."

"Another thing an instructor can do to be enthusiastic is giving personal examples, from his or her own experiences."

"I will now show you two videotaped lectures of the same instructor lecturing on the same topic. As before, the first videotape demonstrates all specific behaviors and therefore, the instructor would be rated at the SA and A end of the rating scale. See if you can spot the specific behaviors as they occur."

[SHOW 3rd VTR: EXAMPLE]

"Did you notice that the instructor varied his voice, the first specific behavior, [pause] and his movements, the second specific behavior, by moving around and moving his arms and hands? He also varied activities by using a chart."

"What about the third specific behavior, presenting subject matter with a personal example? What was the personal example used? It was about a worker who was not working up to standard."

"Now let's watch the same instructor but this time he doesn't do any of the specific behaviors. See if you can determine the difference those specific behaviors make in the instructor's enthusiasm."

[SHOW 3rd VTR: NON-EXAMPLE]

"Did you ever have an instructor like that? In this case, he would be rated toward the SD and D end of the scale."

Summary

"You have just finished viewing examples of an instructor demonstrating as well as not demonstrating the specific behaviors related to each of the 3 general behaviors. Instructors who demonstrate all the specific behaviors would be rated toward the SA and A end of the scale. Instructors who didn't perform any of the specific behaviors would be rated at the SD and D end of the scale."

Practice Observing and Rating with Feedback

Overview

"You will now be shown a series of five videotape scenes each about two minutes long, of different instructors and different topics. These scenes were recorded from live lectures."

[HAND OUT RATING PACKETS]

"You should now have in front of you the rating forms you will be using to rate the instructors in the scenes. Please do not open them until

we come to that part. You will be using them in a little while; so for now, please keep them face down."

First Scene

"For the first instructor, I want you to pay attention to the content; [pause] there'll be questions on it [pause] and the specific behaviors of the instructor; in particular, the specific behaviors we just went over."

"At the end of the scene, you can open the packet in front of you to the first rating form. Here is an example of the rating form. You will answer 3 questions on the content of the lecture. These questions will be True and False or Multiple Choice. Then you will write a check mark in the appropriate blank indicating whether or not (i.e., Yes or No) the specific behaviors occurred. After each set of specific behaviors will be a statement about the general behavior related to them. You are to rate the general behavior on a scale from Strongly Agree to Strongly Disagree by circling the letter indicating the extent to which you agree with that statement."

"Remember, when rating on the general behaviors, consider the occurrences or non-occurrences of the related specific behaviors. After completing the rating form, you will be given feedback on what a group of experienced raters gave as responses. These judges consist of staff members at the Learning & Evaluation Service department, here on campus."

"The instructor in the first scene is lecturing on the principles of positive reinforcement."

[SHOW SCENE NO. 1]

"Now you can turn to the first rating sheet which has 'SCENE NO. 1' at the top and complete it."

[TURN OFF VTR UNTIL EVERYONE IS FINISHED OR 3 MINUTES]

"Now, this is what the group of judges gave as responses to the items on the rating form. You will notice that these judges are not always in agreement with each other. That's because they have different perspectives of instruction. So, they may have been looking at the specific behaviors differently. You may find your ratings different from the judges. That's all right. We just want you to understand why so that next time you will improve your ratings. By the way, this feedback is strictly for training and not to evaluate you."

[GIVE FEEDBACK--SEE "FEEDBACK ON VIDEOTAPE 1" SHEET]

"Please use the bookmark provided and place it behind the first rating sheet; then close your rating packet."

SCENE NO. 2

"You will now be shown a two-minute scene of another instructor. Again, you are to observe the instructor for both content of the lecture and the performance of specific behaviors. After this scene, you will again have to answer questions about the content of the lecture. This time,

however, you will rate the instructor on the 3 general behaviors but not on the specific behaviors. Instead, after rating a general behavior, you are to list the specific behaviors which occurred or did not occur that caused you to rate the way you did. You can list the same specific behaviors we have gone over as well as any other specific behavior related to the general behavior."

"Again, do not discuss your ratings with anyone. And, please do not look back at your previous rating. After rating the instructor, you will be given feedback."

"The topic of this lecture if "The Parents' Authority over their Children."

[SHOW SCENE NO. 2]

"Now you can open your rating packet to the second rating sheet, where the bookmark should be. You can answer the questions now."

[TURN OFF VTR]

[WAIT THREE MINUTES or UNTIL EVERYONE IS FINISHED]

"this is what the group of judges gave as responses to the questions and items on the rating form. Again, these judges are not always in agreement with each other."

[GIVE FEEDBACK--SEE "FEEDBACK ON VIDEOTAPE 2" SHEET]

"Please place the bookmark behind that rating sheet and close your rating packet."

SCENE NO. 3

"The remaining three scenes will also be of different instructors lecturing on different topics. I want you to observe each one as before. But the rating sheets will be different from the 2 you have already used. They will all have 3 questions on the content of the lecture and ratings of the same 3 general behaviors. After each rating, you will be given feedback."

"The next scene is on 'A Total Stimulus Deprivation Experiment in Psychology'."

[SHOW SCENE NO. 3]

"You can rate the instructor now. Remember, there should not be any discussion of the ratings."

[TURN OFF VTR]

[WAIT 1 1/2 MINUTES or UNTIL EVERYONE IS FINISHED]

"This is what the group of judges gave as responses to the questions and items on the rating form."

[GIVE FEEDBACK--SEE "FEEDBACK ON VIDEOTAPE 3" SHEET]

"Place the bookmark behind the rating sheet and close the rating packet."

SCENE NO. 4

"The next scene is on "Sanitary Engineering."

[SHOW SCENE NO. 4]

"You can rate the instructor now."

[TURN OFF VTR]

[WAIT 1 1/2 MINUTES or UNTIL EVERYONE IS FINISHED]

[TURN VTR ON]

"This is what the group of judges gave as responses to the content questions and items on the rating form."

[GIVE FEEDBACK--SEE "FEEDBACK ON VIDEOTAPE 4" SHEET]

"Place the bookmark behind the rating sheet and close the rating packet."

SCENE NO. 5

"The last scene is on "English Composition."

[SHOW SCENE NO. 5]

"You can rate the instructor now."

[TURN OFF VTR]

[WAIT 1 1/2 MINUTES or UNTIL EVERYONE IS FINISHED]

[TURN VTR ON]

"This is what the group of judges gave as responses to the questions and items on the rating form. Let me remind you that the judges may not always be in agreement with each other."

[GIVE FEEDBACK--SEE "FEEDBACK ON VIDEOTAPE 5" SHEET]

"Please close your rating packet and place it face down."

Concluding Comments

"Today's session focused on three general instructor behaviors and at the same time attempted to develop in you a pattern of basing your ratings of general behaviors on observations of specific behaviors. Hopefully, after today's practice and feedback using this process, you will be able to rate your instructor with greater accuracy."

On April 19, a week from this coming Wed., at the beginning of class you will be asked to rate your instructor. A month from now (May 12) you will rate your instructor again. It is important that you be at class to complete these ratings. If you should have any problems which prevent you from being there, I want you to contact Mark Spool at the phone number listed in the handout you will get. We can then make arrangements with you to get your rating some other time. It is also important that you attend the next three classes so you can observe your instructor. Therefore, attending class regularly and rating your instructor on April 19 and May 12 are important for you to receive the full amount of your extra credit."

"One last comment. When you rate your instructor, there will be a place on the answer sheet for an identification number. You will be reminded of this, but we want you to use your student number as an identification. This way we can determine that you met all the requirements for the full amount of extra credit. No one will ever see your ratings, so you can rest assured that your ratings will be confidential."

"I will give you a handout which will remind you when you will be rating your instructor. The handout will also remind you to attend class regularly so that you can observe your instructor and it will also remind you to avoid discussing tonight's session with other classmates."

"Thank you for your cooperation."

[TURN OFF VTR]

[GIVE HANDOUTS TO Ss AS THEY LEAVE]

[COLLECT RATING PACKETS]

FEEDBACK ON SCENE 1:
"Principles of Positive Reinforcement"

Content: [SHOW ANSWERS ON OVERHEAD]

1: "The answer to the first content question is c."
2: "The answer to the second question is True."
3: "The answer to the third question is True."

Specific and General Behaviors:	Yes	No
A_1: "All the judges reported that the instructor did state why the material is being presented by stating the importance of positive reinforcement."	5	
A_2: "On this behavior, the judges were split. Those who said no said so because the instructor did not say something to make it relevant, even though the content is inherently relevant."	2	3
* : "The judges agreed that the instructor made the subject matter relevant; 3 rated SA and 2 A."	3-SA 2-A	
B_1: "Everyone agreed that the instructor did not refer to an outline. There was an outline behind him but the instructor didn't refer to it at all."		5
B_2: "The instructor did give examples of positive reinforcement. Therefore, all the judges said yes."	5	
B_3: "The instructor asked several rhetorical questions, like What would happen if ___?, which the judges felt would get the attention of the students."	5	
* : "The judges generally agreed that the instructor helped the students keep their attention on the subject matter (3 said SA and 2 A)."	3-SA 2-A	
C_1: "The instructor definitely varied his voice."	5	
C_2: "He also moved around."	4	1
C_3: "However, he did not give any personal examples."	1	4
* : "The judges generally agreed the instructor was enthusiastic. Three rated SA and 2 agree."	3-SA 2-A	

FEEDBACK ON SCENE 2:
"The Parent's Authority over Their Children"

Content:

1: "The answer to the first question is b."
2: "The answer to the second question is False."
3: "The answer to the third question is a."

General Behaviors:

				5	
1. Made the subject matter relevant	SA	A	N	D	SD

"The judges all rated Disagree. Even though the subject matter may appear relevant to you, the fact is the instructor did not directly:

a. State why the material was presented or its importance.
b. State how the content related to the students' interests, background or activities."

	2	2	1		
2. Helped keep your attention on the subject matter.	SA	A	N	D	SD

"The judges were mixed on this one. Their average rating was Agree; however, 2 rated SA and one rated N. What affected their ratings were the following specific behaviors:

a. The instructor did not show where each point in the lecture fit into an outline.
b. But, he did give examples, and
c. he did start off with a powerful statement "The parents' authority are, in the child's life, undermined every time by . . ."

	4	1			
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

"Almost all the judges rated this statement SA. The instructor:

a. varied his voice
b. varied his movement, and
c. gave a personal example from his own experience."

FEEDBACK ON SCENE 3:
"A Total Stimulus Deprivation Experiment in Psychology"

Content:

1: "The answer to the first content question is True."
2: "The answer to the second question is b."
3: "The answer to the third question is False."

General Behaviors:

	1		2	2	
1. Made the subject matter relevant.	SA	A	N	D	SD

"Here the judges were split between N and D. Actually, the instructor neither stated why the material was important nor how it related to the students' background or interests. But the subject matter was considered 'inherently' interesting, and therefore some judges put down N."

	3	2			
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD

"The judges generally agreed that the instructor helped keep their attention on the subject matter. They were split between rating SA (3 did) and A (2 did). Their ratings were partially based on the fact that the instructor did give examples and did grab their attention in the beginning with a powerful statement about being totally deprived of all senses, but the instructor did not refer to an outline at all."

	5				
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

"All the judges Strongly Agreed to this statement. Among other things, the instructor varied his voice and movement as well as gave a personal example."

FEEDBACK ON SCENE 4:
"Sanitary Engineering"

Content:

1: "The answer to the first content question is a."
2: "The answer to the second question is c."
3: "The answer to the third question is d."

General Behaviors:

	3	2			
1. Made the subject matter relevant.	SA	A	N	D	SD

"The judges generally agreed to this statement (3 said SA and 2 said A). They said the instructor stated why the topic was important and how the content related to the students."

	4	1			
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD

"Almost all the judges rated this statement Strongly Agree. The instructor gave examples, he grabbed the students' attention from the very beginning and he referred to an outline by saying where they were at along various points in the lecture."

	3	2			
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

"The judges were split between Strongly Agree (3) and Agree (2). All judges said the instructor varied his voice and his movement but they were split on whether or not he presented a personal example.

FEEDBACK ON SCENE 5:
"English Composition"

Content:

1: "The answer to the first content question is b."
2: "The answer to the second question is a."
3: "The answer to the third question is d."

General Behaviors:

	1 SA	2 A	1 N	1 D	SD
1. Made the subject matter relevant.					

"The average rating of the judges was Agree but the judges were still somewhat split on this item. One rated SA, 2 A, 1 N and 1 D. All judges felt the instructor stated why the material was being presented (i.e., its importance) but the judges did not agree with whether or not the instructor stated how the content related to the students' interests, background or activities. Some judges said yes but others said no (i.e., even though the instructor alluded to the relationship with students, he did not state it specifically or directly)."

	1 SA	3 A	1 N	D	SD
2. Helped you keep your attention on the subject					

"The judges generally agreed to this item (3 did). The judges based their ratings on the behaviors of the instructor, such as showing where each point fits into an outline and stating at least one meaningful example or illustration for each major point. The judges, however, were split between whether or not the instructor made a statement to grab students' attention."

	SA	3 A	N	2 D	SD
3. Was enthusiastic when lecturing.					

"The judges were split on rating this item. Three judges rated Agree and 2 rated Disagree. The disagreement between judges on this general behavior is reflected by their disagreement on the occurrence or not of the specific behaviors. The judges were split on agreeing with whether or not the instructor varied his voice and his movement. The judges did agree, however, that the instructor did not present the subject matter with a personal example. More than likely, there were other specific behaviors the instructor did or did not do which affected the judges' ratings."

Improving Ratings of Instruction

The purpose of today's session is to improve student ratings of instruction, like the ones you fill out at the end of the term. Most of the session is on videotape and will last about one hour.

After an introduction, you will learn about four common types of rating errors. These errors should be avoided when rating instructors. You will then watch on videotape an instructor demonstrating three general instructor behaviors characteristic of instructors:

1. Making the subject matter relevant to students
2. Helping students keep their attention on the subject matter
3. Being enthusiastic when lecturing

You will also get a chance to practice rating five different instructors on these general behaviors. The purpose of these practice ratings is to help you improve your ratings, not to evaluate you. You will get some feedback on how a group of experienced observers rated the instructors. Then, you will review your ratings with the common types of rating errors in mind.

To maintain independence among participants, it will be necessary to avoid discussion during today's session, particularly during the practice ratings and feedback.

Next week on Wednesday, April 19, at the end of class you will be asked to rate your instructor. You will also rate your instructor in about a month, on Friday, May 12.

At the end of today's session I will give you a handout with these dates as a reminder.

Your cooperation is greatly appreciated.

SCRIPT FOR RATER ERROR TRAINING PROGRAM

Introduction

Attention Getter

"Have you ever been asked to rate your instructor on something general like 'organization' and then ask yourself, 'I wonder what they mean by that?' [pause] Then, you find out that someone else rated the instructor high while you rated him low."

"It is not uncommon for students to disagree with each other when rating instructors and as a result, instructors have a hard time interpreting their ratings. Which students should they believe?"

Purpose

"It has been said that the quality of student ratings of instruction depend upon the ability of students to accurately observe and rate their instructors. If students can be trained to observe their instructor, then the quality of their ratings should improve."

"We are investigating whether or not this is true. What you will learn from this training session should really help you rate your instructors. This, then, should really have an impact on how instructors look at their ratings."

"Today I will train you in how to accurately observe and rate three (3) general instructor behaviors."

"After today's session, I'll want all of you to pay particular attention to similar teaching behaviors on the part of your instructor back in

your classroom. Next week and again in four weeks you will rate your instructor."

"I'll tell you more about this later, but you should be thinking about it while you go through the training."

Ground Rules

We're going to be running this training during this week with other students in your class. Therefore, we ask you not to discuss tonight's session with other students in your class because that might influence or bias them."

"Similarly, because each of you will be rating your instructor one week and four weeks from now, please do not discuss tonight's session with others within this group during and after tonight's session."

"To summarize, we request that you do the following things":

1. Learn what is taught today as best you can.
2. Avoid discussing the content of today's session with any students in this group and other students in the class.
3. Observe your instructor for the next several weeks--this means you should attend class regularly.
4. Come to class on Wed., April 19, to rate your instructor.
5. Rate your instructor again in class on Friday, May 12.

Overview

"Today's session will last about an hour. I will first lecture on 4 common types of rating errors. I will also review with you 3 general instructor behaviors that are important":

No. 1: Making the subject matter relevant.

No. 2: Helping you keep your attention on the subject matter.

No. 3: Being enthusiastic when lecturing.

"To better understand these general behaviors, you will see short examples of each. Then you will watch a series of 5 short scenes of different instructors and practice rating these instructors on the 3 general behaviors. I will give you feedback after each practice rating on how a group of experienced observers rated that instructor. Finally, you will go over your ratings and determine if you are making one of the common rating errors. We are not interested in evaluating you. We just want you to be aware of the way you rate and to improve it."

Common Rating Errors

"Did you ever get into a discussion with students about your instructor after rating him and found the following situation?"

A: "Boy, that instructor is great. He's so exciting and interesting. Everything he does I rated him high!"

B: "I don't know. I don't think anyone should be rated that high. In fact, I don't believe anyone could be so good or so bad as to deserve any extreme rating. I always mark them near the middle."

"Do you know anyone like those students? Could you be like one of them?"

"These two people are not atypical. In fact, it is because people have these 'styles' or ways of rating others that errors in ratings exist.

If the individual responds in one of the previous ways, regardless of what an instructor does, then he or she is likely to make errors in rating."

"There are four common types of rating errors:

1. Central Tendency
2. Leniency
3. Strictness
4. Halo

I will describe and give an example of each."

Central Tendency

"Raters who commit the 'central tendency' error are those who are reluctant to give high or low ratings. Instead, they tend to continuously use the center or average point on a rating scale even though large differences exist in the behavior of the person being rated."

"Every student is familiar with a similar problem involving the assignment of course grades. There are some instructors who tend to give mostly 2.0s & 2.5s, with very few students getting 3.5s and 4.0s and equally few students getting 1.0s."

"The distribution of ratings of 5 individuals by one rater who makes the central tendency error look like this:

The instructor was organized.

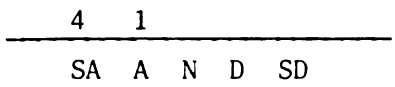
	1	3	1	
SA	A	N	D	SD

"These individuals may in fact be average, but chances are slim that there are no outstanding or inferior people. Therefore, this rater more than likely committed the central tendency error."

Leniency

"Some raters tend to concentrate their ratings toward the upper end of the scale for everyone they rate. These raters make the leniency error. In this case, they only say 'good' things about everyone. You are probably familiar with some instructors who give only 4.0s and 3.5s. But, you know also that some students deserve less than that."

"The distribution of ratings of a rater who commits the leniency error when rating 5 individuals looks like this:
The instructor was organized:



"It is highly unlikely that that many people are that good."

Strictness

"Other raters tend to concentrate their ratings toward the lower end of the rating scale. These raters commit the strictness error. This error is the opposite of the leniency error. Unfortunately, some of you have probably had instructors like this, who give only low grades."

"The distribution of ratings of a rater who makes the strictness error looks like this":

The instructor was organized:

		1		4
SA	A	N	D	SD

"It is also unlikely that that many people are that bad."

Halo

"Halo error refers to the tendency to rate an individual either high or low on many behaviors because of one behavior the rater feels is outstandingly good or bad."

"This can easily happen in student ratings of instructors. If the student feels the instructor is very exciting and interesting and therefore rates the instructor high on everything, then the student is committing the halo error. The student is also committing the halo error if he feels the instructor is completely disorganized and therefore rates the instructor low on everything."

"Equally as problematic is the rater who gives very variable and inconsistent ratings without regard to the actual behaviors of the person being rated. Here the rater gives the false impression that care has been taken in rating the ratee."

"The distribution of ratings of a rater who makes the halo error when rating one individual on 3 behaviors looks like this:

The instructor:

dressed neatly X_____

knew the material X_____

was organized X_____

SA A N D SD

or

dressed neatly _____X

knew the material _____X

was organized _____X

SA A N D SD

"Now I would like you to look at the following distributions by a rater and see if you can determine the error represented by each distribution."

[For each distribution, ask "Which error is this rater making?"]

The instructor is exciting.

[Leniency]

5 1

SA A N D SD

[Central Tendency]

1 5

SA A N D SD

[Strictness]

2 4

SA A N D SD

The I is exciting. X [Halo (also Leniency)]

SA A N D SD

The I is organized. X

SA A N D SD

The I gave fair exams.

X
SA A N D SD

The I started class on time.

X
SA A N D SD

"In summary, there are four (4) types of rating errors we will be concerned about":

1. Central Tendency
2. Leniency
3. Strictness
4. Halo

"Any student is susceptible to any of these rating errors and may not even know it. From today's session we hope that you will be able to determine if you tend to rate in any of the ways mentioned."

"When rating, consider possible rating errors. Let me mention, here, however, that an instructor can certainly be rated SA or SD or even N. To rate an instructor as such does not mean that you are committing a rating error. It is only an error if you rate all instructors in the same way--regardless of their actual performance."

"Later on you will be rating instructors on videotape. We hope that from this practice rating you will discover the degree to which you are prone to rating errors."

Explanation of the Rating Scale

"The rating format for the 3 general behaviors you will be rating will look like this":

The instructor:

1. Made the subject matter relevant.	SA	A	N	D	SD
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
3. Was enthusiastic when lecturing.	SA	A	N	S	SD

"The letters for the rating scale stand for the following:

SA = if you strongly agree with the statement

A = if you agree with the statement

N = if you neither agree nor disagree

D = if you disagree with the statement

SD = if you strongly disagree with the statement

"When you read the statement about the instructor, you are to respond by circling the letter representing the extent [pause] to which you agree with it."

"To help you get a better understanding of the three (3) general instructor behaviors, you are going to see examples of an instructor demonstrating each. It is important to understand that these general behaviors are not totally separate. For example, an instructor being enthusiastic may help you keep your attention on the subject matter. Yet, these two general behaviors are not quite the same--treat them differently."

"For each general behavior, you will first see a videotape of an instructor illustrating the general behavior. In this case, you would rate the instructor at the SA and A end of the rating scale. Then you will see the same instructor lecturing such that you would rate him at the SD and D end of the scale."

"For each general behavior, see if you can spot the difference between the instructor at the SA end of the scale and the same instructor at the SD end of the scale."

Explanation and Demonstration of Three General Instructor Behaviors

"These are the 3 general instructor behaviors you will be rating the instructors on in the last part of today's session."

The instructor:

No. 1: Made the subject matter relevant.

No. 2: Helped you keep your attention on the subject matter.

No. 3: Was enthusiastic when lecturing.

"I will now show you two videotaped lectures by the same instructor on the topic of 'motivation in organizations.' The instructor is lecturing to a group of undergraduate students in management. In the first videotape, the instructor is making the subject matter relevant. Therefore, he would be rated toward the SA and A end of the rating scale."

[SHOW 1st VTR: EXAMPLE]

"The instructor you just watched made the subject matter relevant. You may have felt he wasn't very exciting. If that were the case, and if you were to rate him SD on making the subject matter relevant because of that without respect to his actual behavior, then you would be committing the halo error."

"Now, let's watch the same instructor, but this time he doesn't make the subject matter relevant."

[SHOW 1st VTR: NON-EXAMPLE]

"In this case he would be rated toward the SD and D end of the rating scale."

"The next two scenes show the same instructor, the first time helping the students keep their attention on the subject matter and the second time he doesn't. Let's watch the first scene."

[SHOW 2nd VTR: EXAMPLE]

"This instructor would be rated toward the SA and A end of the rating scale. Now for the second scene."

[SHOW 2nd VTR: NON-EXAMPLE]

"This time the instructor would be rated toward the SD & D end of the scale."

"The third general instructor behavior, that is, 'Was enthusiastic when lecturing,' is more noticeable to students. See if you can spot the difference in the instructor in the following two scenes."

[SHOW 3rd VTR: EXAMPLE]

[SHOW 3rd VTR: NON-EXAMPLE]

"Have you ever had an instructor like either one of these? Remember, however, just because the instructor is or is not enthusiastic does not mean he is or does other things, like make the subject matter relevant."

Summary

"You have just finished viewing examples of an instructor demonstrating the 3 general instructors behaviors as they would be rated toward both ends of the rating scale."

Practice Observing and Rating with Feedback

Overview

"You will not be shown a series of five short scenes, each about 2 minutes long, of different instructors and different topics. These scenes were recorded from live lecturs."

[HAND OUT RATING PACKETS]

"You should now have in front of you the rating forms you will be using to rate the instructors in the scenes. Please do not open them until we come to that part. You will be using them soon, so for now, please keep them face down."

"The rating forms have two parts to them: 3 content questions about the lecture and the 3 general behaviors you are to rate the instructor on. Therefore, for all scenes, I want you to pay attention to the content of the lecture [pause] and the general behaviors of the instructor."

First Scene

"At the end of the first scene, you can open the packed in front of you to the first rating form. You will answer the 3 questions on the content of the lecture. They will be either True and False or Multiple

Choice. Then you will rate the 3 general behaviors by circling the letter which indicates the extent to which the general behavior describes that instructor. After rating, I will give you feedback on what a group of experienced raters or judges gave as responses. These experienced judges consist of staff members at the Learning & Evaluation Service department, here on campus."

"The instructor in the first scene is lecturing on the Principles of Positive Reinforcement."

[SHOW SCENE NO. 1]

"Now you can turn to the first rating sheet which says 'Scene No. 1' at the top and complete it. Remember that there should be no discussion of the ratings with anyone."

[WAIT 2 MINUTES or UNTIL EVERYONE IS FINISHED]

"Now, this is what the group of judges gave as responses to the items on the rating form. You will notice that these judges are not always in agreement with each other. That's because they have different perspectives of instruction. You may find your ratings different from the judges. That's all right. We just want you to understand why so that next time you will improve in your ratings. This feedback is strictly for training and not to evaluate you."

[GIVE FEEDBACK--SEE "FEEDBACK ON SCENE 1" SHEET]

"Please use the bookmark provided and insert it behind the first rating sheet; then close your rating packet."

Scene 2

"The next scene is of an instructor lecturing on 'The Parents' Authority over their Children.'"

[SHOW SCENE NO. 2]

"Now you can open your rating packet to the second rating sheet, where the bookmark should be. You can answer the questions now."

[TURN OFF VTR]

[WAIT 2 MIN. or UNTIL EVERYONE IS FINISHED]

"This is what the group of judges gave as responses to the questions and items on the rating form. Again, these judges are not always in agreement with each other."

[GIVE FEEDBACK--SEE "FEEDBACK ON SCENE 2" SHEET]

"Please place the bookmark behind that rating sheet and close your rating packet."

"The next scene is on 'A Total Stimulus Deprivation Experiment in Psychology.'"

[SHOW SCENE NO. 3]

"You can rate the instructor now."

[TURN OFF VTR]

[WAIT 1 1/2 MINUTES or UNTIL EVERYONE IS FINISHED]

"This is what the group of judges gave as responses to the questions and items on the rating sheet for Scene No. 3."

[GIVE FEEDBACK--SEE "FEEDBACK ON SCENE 3" SHEET]

"Place the bookmark behind the rating sheet and close the rating packet."

Scene No. 4

"The next scene is on 'Sanitary Engineering.'"

[SHOW SCENE NO. 4]

"You can rate the instructor now."

[TURN OFF VTR]

[WAIT 1 1/2 MINUTES or UNTIL EVERYONE IS FINISHED]

"This is what the group of judges gave as responses to the content questions and items on the rating sheet for the fourth scene."

[GIVE FEEDBACK--SEE "FEEDBACK ON SCENE 4" SHEET]

"Place the bookmark behind the rating sheet and close the rating packet."

Scene No. 5

"The last scene is on 'English Composition.'"

[SHOW VIGNETTE NO. 5]

"You can rate the instructor now."

[TURN OFF VTR]

[WAIT 1 1/2 MINUTES or UNTIL EVERYONE IS FINISHED]

"This is what the group of judges gave as responses to the questions and items on the rating form."

[GIVE FEEDBACK--SEE "FEEDBACK ON SCENE 5" SHEET]

"These ratings tend to show that the judges have different perspectives of instructors and their instruction."

"Please close your packets."

Review of Ratings vis-a-vis Rating Errors Discussed

"I would now like to show you in detail some on the judges' ratings with regard to the rating errors discussed earlier."

Leniency Error

[SHOW ON CHART]

Judge A:

		4		1	
1. Made the subject matter relevant.	SA	A	N	D	SD
		4	1		
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
		4	1		
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

Other Judges:

	SA	A	N	D	SD
1. Made the subject matter relevant.	5	6	2	7	
2. Helped you keep your attention on the subject matter.	9	9	2		
3. Was enthusiastic when lecturing.	11	7		2	

"Here are the ratings of one judge in comparison to the other judges. Compare the two distributions. Think about what you notice of Judge A. [pause] He tends to rate toward the SA end of the scale. He may not be completely wrong, however, because some of the other judges agree. But not all agree. Therefore, think of the kind of rating error he could possibly be making. [pause] It's leniency error. [pause] This rater tends to rate toward the upper extreme of the scale on all behaviors."

Halo and Strictness Error

[SHOW ON CHART]

Judge A:

	SA	A	N	D	SD
1. Made the subject matter relevant.				X	
2. Helped you keep your attention on the subject matter.				X	
3. Was enthusiastic when lecturing.				X	

Judge B:

		SA	A	N	D	SD
1.	Made the subject matter relevant.		X			
2.	Helped you keep your attention on the subject matter.		X			
3.	Was enthusiastic when lecturing.				X	

"Now, here are the ratings of the same instructor by two different judges. The first judge rated the instructor Disagree on the all 3 general behaviors. The other judge rated Disagree only on the third general behavior and Agree on the first 2 general behaviors."

"There are 2 possible rating errors Judge A may be making. If you think that the first judge was influenced by his rating on the 3rd general behavior (i.e., the instructor's lack of enthusiasm)? The first judge, therefore, may be committing which rating error? [pause] Halo error."

"However, if Judge A rated everyone low on all behaviors, he mat be committing which rating error? [pause] Strictness."

"Here's the distribution of all judges' ratings on all 5 instructors.

[SHOW ON CHART--SEE END OF SCRIPT]

"None of the judges committed the central tendency error (i.e., rating all Ns) or the strictness error (i.e., rating all SDs).

"Now I want you to turn to the last sheet in your rating packet, mark the number of times you rated the instructors at each point on the

scale. Do this separately for each of the 3 general behaviors. An example is provided at the top of that sheet. Then compare your distribution of ratings with that of the judges' and see if you might be prone to one of the rating errors. Don't worry if you think you are. This exercise is just to help you become aware of likely rating errors and to improve your ratings."

"Take about 5 minutes to do this. Do not compare or discuss your ratings with anyone else."

[WAIT 5 MINUTES--SHOW JUDGES' COMPOSITE RATINGS ON SCREEN]

"Please turn your rating packet face down."

Concluding Comments

"Today's session focused on 3 general instructor behaviors and at the same time attempted to help you become aware of possible rating errors you may be likely to make. Hopefully, after today's practice and feedback, and knowing probable rating errors, you will be able to rate your instructor with greater accuracy."

"On April 19, a week from this coming Wed., at the beginning of class you will be asked to rate your instructor. A month from now (May 12) you will rate your instructor again. It is important that you be at class to complete these ratings. If you should have any problems which prevent you from being there, I want you to call Mark Spool at the phone number listed in the handout you will be given soon. We can then make arrangements with you to get your rating some other time. It

is also important that you attend the next three classes so you can observe your instructor. Therefore, attending class regularly and rating your instructor on April 19 and May 12 are important for you to receive the full amount of your extra credit."

"One last comment. When you rate your instructor, there will be a place on the answer sheet for your student number. You will be reminded of this. This way we can determine that you met all the requirements for the extra credit. No one, however, will ever see your ratings, so you can rest assured that your ratings will be confidential."

"You will get a handout which will remind you when you will be rating your instructor. The handout will also remind you to attend class regularly so that you can observe your instructor and it will also remind you to avoid discussing today's session with other classmates."

"As you leave, turn in your rating packet."

"Thank you for your cooperation."

[TURN OFF VTR]

[GIVE HANDOUTS TO Ss AS THEY LEAVE]

[COLLECT RATING PACKETS]

Distribution of All Judges'
Ratings on all 5 Instructors

1.	Made the subject matter relevant.	9	6	3	7	
		SA	A	N	D	SD
2.	Helped you keep your attention on the subject matter.	13	10	2		
		SA	A	N	D	SD
3.	Was enthusiastic when lecturing.	15	8		2	
		SA	A	N	D	SD

FEEDBACK ON SCENE 1:
"Principles of Positive Reinforcement"

Content:

1. "The answers to the first content question is c."
2. "The answer to the second question is True."
3. "The answer to the third question is True."

General Behaviors:

		3	2			
1.	Made the subject matter relevant.	SA	A	N	D	SD
2.	Helped students keep their attention on the subject matter.	SA	A	N	D	SD
3.	Was enthusiastic when lecturing.	SA	A	N	D	SD

FEEDBACK ON SCENE 2:
"The Parents' Authority over their Children"

Content:

1. "The answer to the first question is b."
2. "The answer to the second question is False."
3. "The answer to the third question is a."

General Behaviors:

1. Made the subject matter relevant.	SA	A	N	5 D	SD
2. Helped students keep their attention on the subject matter.	2 SA	2 A	1 N	D	SD
3. Was enthusiastic when lecturing.	4 SA	1 A	N	D	SD

FEEDBACK ON SCENE 3:
"A Total Stimulus Deprivation Experiment in Psychology"

Content:

1. "The answer to the first content question is True."
2. "The answer to the second question is b."
3. "The answer to the third question is False."

General Behaviors:

1.	Made the subject matter relevant.	1 SA	A	2 N	2 D	SD
2.	Helped students keep their attention on the subject matter.	3 SA	2 A	N	D	SD
3.	Was enthusiastic when lecturing.	5 SA	A	N	D	SD

FEEDBACK ON SCENE 4:
"Sanitary Engineering"

Content:

1. "The answer to the first content question is a."
2. "The answer to the second question is c."
3. "The answer to the third question is d."

General Behavior:

1.	Made the subject matter relevant.	3 SA	2 A	N	D	SD
2.	Helped students keep their attention on the subject matter.	4 SA	1 A	N	D	SD
3.	Was enthusiastic when lecturing.	3 SA	2 A	N	D	SD

FEEDBACK ON SCENE 5:
"English Composition"

Content:

1. "The answer to the first content question is b."
2. "The answer to the second question is a."
3. "The answer to the third question is d."

General Behaviors:

The instructor:

1.	Made the subject matter relevant.	1 SA	2 A	1 N	1 D	SD
2.	Helped students keep their attention on the subject matter.	1 SA	3 A	1 N	D	SD
3.	Was enthusiastic when lecturing.	SA	3 A	N	2 D	SD

Reminder Sheet

We're going to be running this training program during this week with other students in your class. Therefore, we ask you not to discuss tonight's session with other students in your class because that might influence or bias them. Similarly, because each of you will be rating your instructor one week and four weeks from now, please do not discuss tonight's session with others within this group after tonight's session.

As a reminder, you will be rating your instructor at the end of class on Wednesday, April 19 and again on Friday, May 12. In order to determine your participation in this study, you must write your student number on the rating sheet. Your instructor, however, will not see these ratings.

At the end of the term you will be debriefed about the study.

Thank you for participating.

Mark Spool

General Behavior No. 1: "Making the subject matter relevant"
Example of instructor toward "Strongly Agree" end of scale (demonstrating specific behaviors).

"In most jobs in industry, the best worker produces two to three times as much as the worst worker. In some jobs there are differences even greater than this. Today's lecture is about one of the major reasons why workers produce at different rates: [pause] motivation [pause]."

"Motivation is certainly not the only factor that causes people to produce at different rates. The performance level of an individual is influenced by many factors, like the worker's ability and the condition of the machines."

"Still, particularly in the case of lower-level jobs where little ability is required, motivation seems to be the single [pause] most important [pause] determinant of performance."

"The study of motivation is of concern to organizations because it can save them millions of dollars. For those who will be managers in industry, you will be dealing with the issue of motivation a lot. You will be developing programs or redesigning jobs so that workers will be motivated to work hard. If you major in psychology or business management, you will be taking courses which will get into the different theories of motivation."

General Behavior No. 1: "Making the subject matter relevant"
Example of instructor toward "Strongly Disagree" end of scale (NOT demonstrating specific behaviors).

"In most jobs in industry, the best worker produces two to three times as much as the worst worker. In some jobs there are differences even greater than this. Today's lecture is about one of the major reasons why workers produce at different rates: [pause] motivation [pause]."

"Motivation is certainly not the only factor that causes people to produce at different rates. The performance level of an individual is influenced by many factors, like the worker's ability and the condition of the machines."

"In today's lecture we will be looking at motivation and how it affects performance. We will also see where the ability level of a worker fits into determining his performance and how it can possibly interact with his motivation level."

General Behavior No. 2: "Helping you keep your attention on the subject matter"

Example of instructor toward "Strongly Agree" end of scale (demonstrating specific behaviors).

"Here's a situation which a hypothetical manager was faced with one time. The workers in his shop were on a pay-incentive system such that when their production exceeded 67 percent of what had been determined to be average or standard production, they got paid more."

"The manager had two people who stood out from the others. One person produced over 150 percent of standard. [pause] Another worker, however produced an average of only 52 percent of standard." [pause]

"Why do you think there was such a wide difference between these two workers? [pause] Was the first motivated more? [pause] Or, was it that the second worker did not have the ability to do the work?"

"Today, as you can see on our outline [REFER TO OUTLINE], we will be talking about how the system of pay can affect a worker's motivation."

"As an example, let's go back to the situation mentioned earlier and consider the reason given by the first worker for his performance. This worker, who produced at 150%, said he was out to make money. He told the manager, 'I keep my bills paid and I don't owe anybody a damn cent.'"

"The view of this worker illustrates that pay is very important in motivating some workers."

OUTLINE

I. Motivated Behavior
II. Individual Needs
III. Design of Jobs & Performance
IV. System of Pay & Motivation
V. Summary

General Behavior No. 2: "Helping you keep your attention on the subject matter"

Example of instructor toward "Strongly Disagree" end of scale (NOT demonstrating specific behaviors).

"Many managers are faced with situations in which a few employees work way above the standard while at the same time a few work way below standard. There are different reasons for this, of course. To some workers, money is very important. To others, it isn't near as important as things like having friends at work, meeting and talking with people, the challenge of the job, and so on. Therefore, it is important to take into consideration a person's needs when assigning him or her to a job."

"There are many ways that jobs can be designed to increase the satisfaction and thus performance of workers. More specifically, there are five ways that jobs can be designed so as to increase the satisfaction and performance of workers. If jobs are designed so that the worker: (1) has some autonomy, (2) finds the work significant, (3) has some identity with the product, (4) has variety in his job and (5) receives feedback about his performance, then workers may become more satisfied and may produce more."

General Behavior No. 3: "Being enthusiastic when lecturing"

Example of instructor toward "Strongly Agree" end of scale (demonstrating specific behaviors).

VARY VOICE

VARY MOVEMENT/ACTIVITY

"As stated in the beginning of this lecture, job performance is influenced by factors other than motivation. One of the most important factors is ability. [PAUSE] No matter how motivated a worker is to perform well, good performance is not possible if he or she lacks the necessary ability."

"Many theorists have suggested that the following equation expresses the relationship of ability and motivation to performance:

Performance = f(Ability x Motivation)

[Performance is a function of ability and motivation.]

"An important implication of this equation is that not all performance problems that occur in organizations are caused by low motivation. Often, particularly in higher-level jobs, performance problems are caused by low ability."

"Let me tell you about something that happened to me once. When I was hired as a manager in a small company, I was put in charge of three other managers, two of whom I hired. I soon noticed that Mr. Jones, the manager I did not hire, was not performing as well as the others. [pause] Mr. Jones has been with the company for 15 years."

"It first occurred to me that he was not motivated. Perhaps I was giving more attention to the other managers. After talking with him about the problem, I thought everything would be all right. But

it wasn't. It was hard for me to realize but the fact was that Mr. Jones, even though he had 15 years experience, didn't have a lot of new management skills which the other two managers had. It was his ability, not his motivation, that was the problem"

General Behavior No. 3: "Being enthusiastic when lecturing"

Example of instructor toward "Strongly Disagree" end of scale (NOT demonstrating specific behaviors).

DO NOT VARY VOICE

DO NOT VARY MOVEMENT/ACTIVITY

"As stated in the beginning of this lecture, job performance is influenced by factors other than motivation. One of the most important factors is ability. No matter how motivated a person is to perform well, good performance is not possible if the person lacks the necessary ability."

"Many theorists have suggested that performance is a function of ability and motivation. An important implication of this is that not all performance problems that occur in organizations are caused by low motivation. Often, particularly in higher-level jobs, performance problems are caused by low ability."

"When diagnosing the performance problem of individuals in organizations, it is crucial to try to find out how much of the problem is due to poor ability and how much of it is due to low motivation. Poor performance caused by low motivation clearly requires different kinds of corrective action from that required by performance caused by low ability."

APPENDIX C

RATING FORMS FOR TRAINING PROGRAMS

APPENDIX C

_____
Your Student Number

OBSERVATION AND RATING SHEETS

--Please Keep Closed Until Instructed to Open--

There are five rating sheets attached. They are identified by a scene number. You will be given instructions on how to use these rating sheets.

There are a couple of important things to keep in mind. We request that you do not look at the rating sheets before the scene is shown. A bookmark is provided for you to place where you will be turning to next. Also, please do not refer to your previous ratings.

Make sure your student number is at the top of this page. Thank you.

SCENE 1

1. According to the instructor, when a student in class acts silly and the class laughs at him, their laughing is a:

   a. negative reinforcer.
   b. neutral reinforcer.
   c. positive reinforcer.
   d. none of the above

2. According to the instructor, when one child hits another, he does it to get certain consequences which he wants or needs.

   T F

3. According to the instructor, positive reinforcement occurs when the consequences of a behavior promote the reoccurrence of that behavior.

   T F

Check the appropriate blank indicating whether (Yes) or not (No) the specific instructor behavior occurred. Also, respond to the statement marked by an * by circling the letter indicating the extent to which you agree with it using the following scale:

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

The instructor:	Yes	No
A. 1. Stated why the material is being presented (e.g., the importance of the topic).	___	___
2. Stated how the content relates to your interests, background or activities.	___	___
* Made the subject matter relevant.	SA A N D SD	
B. 1. Made a statement to grab students' attention (e.g., a puzzling question, a contradictory or powerful statement, etc.).	___	___
2. Showed where each point fits into an outline, especially as he comes to it.	___	___

		Yes	No
	3. Stated at least one meaningful example or illustration for each major point.	___	___
	* Helped you keep your attention on the subject matter.	SA A N D SD	
C.	1. Varied voice (volume, speed, pitch).	___	___
	2. Varied movement/activity; did not just remain still.	___	___
	3. Presented subject matter with personal examples from his/her own experiences.	___	___
	* Was enthusiastic when lecturing.	SA A N D SD	

SCENE 2

1. According to the instructor, parents' authority of their child's life is undermined by:

   a. the mother spending too much time with the child.
   b. nurses and doctors taking the child away from his parents after birth.
   c. Dr. Spock.
   d. a mother's physical exhaustion.

2. The instructor stated that it has not yet been proven that the mother is more important than the father in a young child's life.

   T F

3. What is one thing this instructor would recommend so that the parents' authority is not undermined?

   a. giving the mother adequate training in child-rearing.
   b. set up a basket and pully system to get the child from the nursery to the mother's room.
   c. have the mother spend a large amount of time (lying-in) with the child.

Respond to the statement according to the extent to which you agree with it using the following scale:

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

The instructor:

1. Made the subject matter relevant. SA A N D SD

   Reasons why (list specific behaviors):

2. Helped you keep your attention on the subject matter. SA A N D SD

   Reasons why (list specific behaviors):

3. Was enthusiastic when lecturing. SA A N D SD

Reasons why (list specific behaviors):

SCENE 3

1. The instructor equated total stimulus deprivation to being "physiologically numb."

   T F

2. Why did the instructor feel that going to the bathroom was "one of the most glorious feelings of his life"?

   a. It was the only thing the experimenter gave him permission to do.
   b. It was one sensation which they had not been able to deaden.
   c. He wasn't sure if he could.

3. The instructor said that in total stimulus deprivation he could not feel his fingers, but he could still see things.

   T F

Respond to the statement according to the extent to which you agree with it using the following scale:

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

The instructor:

1. Made the subject matter relevant.	SA	A	N	D	SD
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

SCENE 4

1. The instructor stated that waste water treatment should not concern the public:

   a. if the sanitary engineer and the urban planner do their jobs right.
   b. because "even a two-year old can do it."
   c. because if is not a topic the public should know about.
   d. if we go back to using commodes.

2. According to the instructor, what was used just before the flush toilet?

   a. the nearest bush
   b. the outhouse
   c. the commode
   d. the chamber pot

3. What did the instructor say will happen if waste water is not properly disposed of?

   a. Man will have to go back to using outhouses.
   b. Life as we know it will no longer exist.
   c. Many sanitary engineers will lose their jobs.
   d. Disease and polluted water will greatly increase.

Respond to the statement according to the extent to which you agree with it using the following scale:

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

The instructor:

1. Made the subject matter relevant.	SA	A	N	D	SD
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

SCENE 5

1. What did the student say was wrong with the first paragraph on the overhead?

   a. It was too long.
   b. There were too many sentences beginning with he.
   c. There were not enough adjectives.
   d. It was too short.

2. What did the instructor say was wrong with the paragraph?

   a. Almost every sentence began with a subject-verb pattern.
   b. Almost every sentence ended with a subject-verb pattern.
   c. Almost every sentence did not have the subject-verb pattern.

3. Why did the instructor say sentence variety was important?

   a. It makes paragraphs more interesting.
   b. It keeps paragraphs from being boring.
   c. It is important for professional writers.
   d. all of the above

Respond to the statement according to the extent to which you agree with it using the following scale:

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

The instructor:

1. Made the subject matter relevant.	SA	A	N	D	SD
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

_____
Your Student Number

OBSERVATION AND RATING SHEETS

--Please Keep Closed Until Instructed to Open--

There are five rating sheets attached. They are identified by a scene number. You will be given instructions on how to use these rating sheets.

There are a couple of important things to keep in mind. We request that you do not look at the rating sheets before the scene is shown. A bookmark is provided for you to place where you will be turning to next. Also, please do not refer to your previous ratings.

Make sure your student number is at the top of this page. Thank you.

SCENE 1

1. According to the instructor, when a student in class acts silly and the class laughs at him, their laughing is a:

   a. negative reinforcer.
   b. neutral reinforcer.
   c. positive reinforcer.
   d. none of the above

2. According to the instructor, when one child hits another, he does it to get certain consequences which he wants or needs.

   T F

3. According to the instructor, positive reinforcement occurs when the consequences of a behavior promote the reoccurrence of that behavior.

   T F

Respond to the statement according to the extent to which you agree with it using the following scale:

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

The instructor:

1. Made the subject matter relevant.	SA	A	N	D	SD
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

SCENE 2

1. According to the instructor, parents' authority of their child's life is undermined by:

   a. the mother spending too much time with the child.
   b. nurses and doctors taking the child away from his parents after birth.
   c. Dr. Spock.
   d. a mother's physical exhaustion.

2. The instructor stated that it has not yet been proven that the mother is more important than the father in a young child's life.

   T F

3. What is one thing this instructor would recommend so that the parents' authority is not undermined?

   a. giving the mother adequate training in child-rearing.
   b. set up a basket and pully system to get the child from the nursery to the mother's room.
   c. have the mother spend a large amount of time (lying-in) with the child.

Respond to the statement according to the extent to which you agree with it using the following scale:

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

The instructor:

1. Made the subject matter relevant.	SA	A	N	D	SD
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

SCENE 3

1. The instructor equated total stimulus deprivation to being "physiologically numb."

   T F

2. Why did the instructor feel that going to the bathroom was "one of the most glorious feelings of his life"?

   a. It was the only thing the experimenter gave him permission to do.
   b. It was one sensation which they had not been able to deaden.
   c. He wasn't sure if he could.

3. The instructor said that in total stimulus deprivation he could not feel his fingers, but he could still see things.

   T F

Respond to the statement according to the extent to which you agree with it using the following scale:

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

The instructor:

1. Made the subject matter relevant.	SA	A	N	D	SD
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

SCENE 4

1. The instructor stated that waste water treatment should not concern the public:

   a. if the sanitary engineer and the urban planner do their jobs right.
   b. because "even a two-year old can do it."
   c. because it is not a topic the public should know about.
   d. if we go back to using commodes.

2. According to the instructor, what was used just before the flush toilet?

   a. the nearest bush
   b. the outhouse
   c. the commode
   d. the chamber pot

3. What did the instructor say will happen if waster water is not properly disposed of?

   a. Man will have to go back to using outhouses.
   b. Life as we know it will no longer exist.
   c. Many sanitary engineers will lose their jobs.
   d. Disease and polluted water will greatly increase.

Respond to the statement according to the extent to which you agree with it using the following scale:

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

The instructor:

1. Made the subject matter relevant.	SA	A	N	D	SD
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

SCENE 5

1. What did the student say was wrong with the first paragraph on the overhead?

   a. It was too long.
   b. There were too many sentences beginning with he.
   c. There were not enough adjectives.
   d. It was too short.

2. What did the instructor say was wrong with the paragraph?

   a. Almost every sentence began with a subject-verb pattern.
   b. Almost every sentence ended with a subject-verb pattern.
   c. Almost every sentence did not have the subject-verb pattern.

3. Why did the instructor say sentence variety was important?

   a. It makes paragraphs more interesting.
   b. It keeps paragraphs from being boring.
   c. It is important for professional writers.
   d. all of the above

Respond to the statement according to the extent to which you agree with it using the following scale:

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

The instructor:

1. Made the subject matter relevant	SA	A	N	D	SD
2. Helped you keep your attention on the subject matter.	SA	A	N	D	SD
3. Was enthusiastic when lecturing.	SA	A	N	D	SD

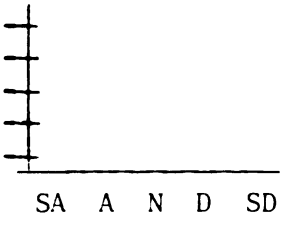
COMPOSITE RATINGS

Example

Ratings on General Behavior No. 1 for:

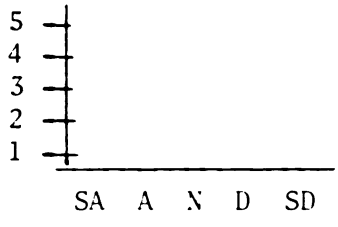
Scene 1 = SA
Scene 2 = SD
Scene 3 = N
Scene 4 = N
Scene 5 = A

Composite Rating =

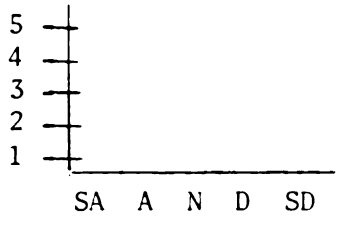


Your ratings (for each General Behavior, for all five instructors):

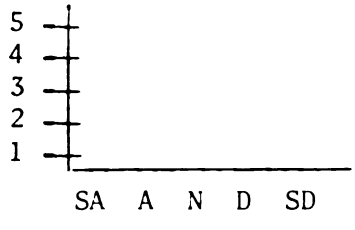
1. Made the subject matter relevant.



2. Helped you keep your attention on the subject matter.



3. Was enthusiastic when lecturing.



APPENDIX D

DEVELOPMENT AND PILOT OF THE "BEHAVIOR OBSERVATION TRAINING PROGRAM"

APPENDIX D

DEVELOPMENT AND PILOT OF THE "BEHAVIOR OBSERVATION TRAINING PROGRAM"

In this appendix, the development and pilot of the Behavior Observation Training Program, including the application of the principles of learning and motivation and the principles of transfer, are discussed.

Development

The results of the ANOVA on the SRI-A pilot rating form suggest that those items containing general instructor behaviors related to 'maintaining attention' are rated with the least amount of interrater agreement among students. It was decided, therefore, to direct training toward observing general and specific instructor behaviors related to 'maintaining attention.'

Examination of those items in both SRI forms revealed three general instructor behaviors: (1) making the subject matter relevant, (2) helping students keep their attention on the subject matter, and (3) being enthusiastic. Following Rosenshine and Furst's (1971) suggestion, a list of specific behaviors was developed for subjects to use when rating the general behaviors. Observing (i.e., "looking for") specific behaviors and reporting them when rating general instructor behaviors, according to Rosenshine and Furst, should improve such ratings. The specific behaviors were derived from information

obtained from faculty and staff employed at the Michigan State University's Learning and Evaluation Service (L&ES), the literature, and the investigator's personal experiences in this area. A list of the items representing specific behaviors related to each of the three general behaviors is shown in the first rating form, found in Appendix C. Because of time limitations on training and ability limitations of subjects, the list was limited to two or three specific behaviors for each general behavior.

The level of a subject coming into training may be different from other subjects. Therefore, to bring all subjects up to the same level of competency, simulated lectures illustrating the three general behaviors at each end of the rating scale (i.e., all specific behaviors present or absent) were videotaped. A male actor (an L&ES faculty member) was obtained for this purpose. The lectures were all on the same topic, "motivation in organizations."

It was determined from the literature review on the principles of learning and motivation that trainees must practice. To provide subjects with realistic practice, videotapes of different instructors presenting live lectures were obtained from L&ES files. A segment from each videotape, meaningful in content and no more than three minutes in length, was selected. These segments will be referred to hereafter as vignettes. Prior to the pilot, there were ten vignettes varying in topic (e.g., psychology, forestry, sociology and sanitary engineering).

Five judges (three L&ES faculty and two L&ES graduate research assistants) experienced in observing instructors rated the vignettes

on: (1) whether or not each specific behavior occurred and (2) each of the three general behaviors from Strongly Agree to Strongly Disagree. Their responses (i.e., the average of judges' ratings and the reasons for their ratings) were used for feedback to the subjects. The vignettes were also sequenced in order from simple to more complex as determined by the variability in judges' ratings.

Three different rating forms were developed so that subjects could practice and receive feedback on observing and rating and also learn how to base their ratings of general behaviors on observations of specific behaviors. To enhance transfer, the rating forms increased in fidelity, that is, reflecting the typical rating form which consists mainly of general behaviors (items) only, and complexity--beginning with cues (i.e., specific behaviors) and ending without those cues. The first rating form followed Rosenshine and Furst's (1971) suggestion that specific behaviors related to a general behavior be rated first. These specific behaviors serve as cues for the rating of the general behavior which follows. The second rating form has fewer cues by requiring subjects to recall and write the specific behaviors after rating the general behavior. The third form, which has no cues, requires subjects to rate the instructor on the general behaviors (items) only, like typical rating forms.

Students in the classroom not only observe the behaviors of their instructor but also listen/observe to learn the content. Therefore, to have high fidelity and an equivalent load, all subjects were required to answer three questions on the content of the vignette. These content questions appeared at the top of the rating sheet.

Application of principles of training

The purpose of this section is to briefly point out the application of the various principles of training to the development and design of the training program. As mentioned in Chapter II, these principles are interactive--one affects the other. Each of the principles will be described briefly. The first eight principles are related to learning and motivation while principles nine through 12 focus on transfer of training. Definitions and explanations of these principles were given in the Review of the Literature (Chapter II).

1. "Meaningfulness" was applied in the introduction by telling the subjects how the training program is relevant to them as students and the benefits of such training.

2. "Prerequisites" was handled by starting everyone from a common entry level through training subjects to recognize clear-cut examples of the specific behaviors as well as showing subjects examples of the general instructor behaviors to be rated.

3. "Modeling" was taken into account during feedback when subjects were told how and why experienced judges rated the instructors. Also, similar to the concept of modeling, examples of the general behaviors were shown to the subjects.

4. "Open communication" was applied by stating the objective of the training and giving an overview/agenda. Because discussion during the practice and feedback part of training was not allowed, the principle of open communication was only partially applied. However, pilot training enabled typical questions to be anticipated and answers were incorporated into the training presentation.

5. "Novelty," through the use of different instructors/vignettes, occurred throughout the training.

6. "Active appropriate practice" was handled through the practice rating part of training, with feedback.

7. "Fading" was applied by proceeding from the first rating form (providing specific behaviors as cues) through the second (requiring recall of specific behaviors) to the remaining rating forms (rating general behaviors only).

8. "Pleasant conditions and consequences" was applied through the use of a comfortable training atmosphere and no public embarrassment for subjects if they made ratings different than the judges'.

9. "General principles" was applied by telling and reminding subjects that ratings of general behaviors should be based upon observations of specific behaviors.

10. "Response availability" was used by sequencing vignettes from simple to more complex.

11. "Identical elements" was approximated by placing an equivalent load upon subjects (by their answering content questions in addition to rating behaviors) and by using a rating form which closely approximates typical student rating of instruction forms.

12. "Performance feedback" was not applied since the experimental design required two observation periods to assess the effectiveness of the training program. Feedback after the first observation might affect the second.

In all, the present training program applied almost all of the principles to enhance learning and motivation and the transfer of training.

Pilot

The training program was piloted by the investigator on three undergraduate students who were enrolled in a course identical to the one that was used in the study. These students were asked by their instructor to participate and do so without extra credit. The pilot lasted three hours due to feedback received about the training.

The training program was presented from a prepared script. At various points in the training program, the trainer stopped and asked questions to facilitate feedback about the training. Feedback, then, was received after: (1) the introduction, (2) the presentation on rating general behaviors, (3) subjects viewed the videotaped examples of the general behaviors, (4) subjects rated the first vignette, (5) subjects rated the second vignette, (6) subjects rated the last vignette and (7) the whole training program. Some of the questions asked to stimulate feedback were: "Tell me what you're supposed to do after this training session?"; "What is coming up in the rest of the training session?"; "Were you bored watching the videotaped examples of the general behaviors?".

Their suggestions varied. Examples are: "Revise the videotape examples of the general behaviors (e.g., show several specific behaviors together demonstrated by one person on the same topic)"; "Provide handouts which include the agenda and 'ground rules' to participants as they come in"; and "Reduce the number of practice vignettes from ten to five--any more would be fatiguing and inefficient." These suggestions and others were incorporated in a revised training program. The script of the final training program, which was videotaped, can be found in Appendix B.

APPENDIX E

DEVELOPMENT OF SRI-A AND SRI-B RATING FORMS

APPENDIX E

DEVELOPMENT OF SRI-A AND SRI-B RATING FORMS

In this appendix the development of the rating forms, their pilot administration and the results of the pilot are discussed. The purpose of piloting the SRI A and B forms was twofold: (1) to reduce the number of items in the forms and (2) to collect some baseline data to use for the development of two training programs.

Development of Pilot Rating Forms

It was desired to limit the rating forms to two areas of instructor behavior. Further, items with at least 83 percent agreement were preferred, with 67 percent agreement being the lowest level acceptable. The 163 items categorized (see Appendix A) with 83 percent or more agreement (see Table A1) fell into three broad areas: Presentation Skills, Organization/Structure and Student Rapport. Table A2 shows the distribution of items across areas and type of item for 100 percent agreement and 83 percent agreement.

Examination of this breakdown revealed that most acceptable items fell within the Presentation Skills area. Further examination of these items indicated that most were related to one of two general instructor behaviors: open communication and maintaining attention. "Open communication" refers to instructor behaviors which link information about the course to students' learning, such as giving

Table A2

Content Areas of Categorized Items

Type of Item		Area of Instructor Behavior											
		Presentation Skills				Organizational/Structure				Student Rapport			
		S-D	S-E	G-D	G-E	S-D	S-E	G-D	G-E	S-D	S-E	G-D	G-E
Level of Agreement	6/6	21	59	9	8	19	2	1	1	9	0	7	7
	5/6	*	4	17	12	*	1	2	2	*	0	4	3

*not calculated

students information about the course, holding discussions, and asking and answering questions. "Maintaining attention" refers to instructor behaviors which actively get students to attend to the subject matter, by making it meaningful (e.g., stating the importance and relevance of the subject matter), by stimulating the students, and by appearing interesting and "enthusiastic."

The breakdown of items into these two general instructor behavior areas is shown in Table A3.

Table A3

Further Breakdown of Items into Content Areas

Types of Items		Area of Instructor Behavior: Open Communication S-D	S-E	G-D	G-E	Maintaining Attention S-D	S-E	G-D	G-E
Level of Agreement	5&6/6	14	4	14	6	14	2	4	10
	4/6	*	1	6	3	*	3	4	2

*not determined (no additional items needed)

The goal was to select ten items (five for the pilot SRI-A form and five for the pilot SRI-B form) in each item category/type and in both areas of instructor behavior. To do this, more items specifically related to S-E and G-E categories under Open Communication and S-E, G-D, and G-E categories under Maintaining Attention were needed. Therefore, 38 additional items were written with the intent of increasing the number of acceptable (at least 67 percent agreement)

items in these areas. The same six judges, using the same instructions (see Appendix A), categorized these items. This resulted in an additional 29 acceptable items. The final breakdown of items is shown in Table A4.

Table A4

Final Breakdown of Items

	Area of Instructor Behavior							
	Open Communication				Maintaining Attention			
Types of Items	S-D	S-E	G-D	G-E	S-D	S-E	G-D	G-E
Number of Acceptable Items	15	11	19	14	19	10	12	15

Within each content (instructor behavior) area-type of item combination, pairs of items similar in wording were identified. Each pair represented one item for the pilot SRI-A form and its "parallel" item for the pilot SRI-B form. Five pairs of items, which represented the widest range of behaviors, within each area were selected. All decisions were subjective and made by the investigator based upon his knowledge and experiences.

In total, there were 40 items (5 items x 4 types of items x 2 content areas) in each pilot rating form. The items in the pilot SRI-A form were randomly ordered. The items in the SRI-B form had the same order as in the SRI-A form.

Procedure and Analysis of Pilot

Ninety-five undergraduate students in an introductory industrial/organizational psychology course, identical to one of the classes in the dissertation, completed the SRI-A (n=49) and SRI-B (n=46). They were given these forms after completing the standard student rating form used at Michigan State University. The instructor asked the students to write comments on the form about the appropriateness and quality of the items used in these rating forms in addition to rating the instructor.

Item means and standard deviations in each subtest (Behavior-Judgment-Content Area combination) were calculated. Also, a 2x2x2 (Behavior x Judgment x Content Area) ANOVA with repeated measures on all three factors was performed on the SRI-A pilot rating form only to test the effects of level of Behavior, level of Judgment and Content Area on interrater agreement among raters. Interrater agreement was calculated by squaring the deviation of a subject's average item score within a subtest from the average item score for the same subtest for all subjects.

Results

The means and standard deviations of the items in the SRI-A and SRI-B pilot forms are shown in Table A5.

Results of the ANOVA with repeated measures are summarized in Table A6. There was a significant difference between the two content areas, $F(1,384) = 7.78$, $p < .005$. The main effect for level of Judgment, however, was not significant. There was a significant difference between the levels of Behavior, $F(1,384) = 5.85$, $p < .016$.

Table A5

Means and Standard Deviations of Pilot

SRI-A and SRI-B Items

	SRI-A		SRI-B		
Item No.	$\bar{X}$	SD	$\bar{X}$	SD	Subtest
1	3.5	.91	3.3	.89	S-E (OC)
2	3.4	1.06	2.9	.98	G-D (MA)
3	2.1	.74	2.9	1.07	S-D (OC)
4	3.4	.84	3.3	1.03	G-E (OC)
5	4.0	.91	3.6	1.20	G-E (MA)
6	2.2	.97	2.8	.90	S-E (MA)
7	2.7	.81	2.4	.78	G-D (OC)
8	3.2	1.03	2.3	.87	G-D (MA)
9	2.1	.81	2.2	.61	S-D (OC)
10	2.6	.82	2.6	.95	S-E (OC)
11	4.2	.86	3.2	1.20	S-D (MA)
12	4.1	.82	3.7	1.04	G-E (MA)
13	2.6	.88	2.3	.81	G-E (OC)
14	4.2	.75	3.8	.99	S-E (MA)
15	3.6	.70	3.2	.92	G-E (OC)
16	2.2	.80	2.3	.77	G-E (OC)
17	3.5	.82	3.9	.95	S-D (OC)
18	3.4	.87	3.2	1.05	G-D (OC)
19	3.1	.96	2.7	1.00	G-D (MA)
20	2.8	.75	2.6	1.02	G-E (OC)
21	3.3	.94	3.9	.83	G-D (OC)
22	2.6	.91	2.3	.99	G-D (MA)
23	4.2	.90	2.0	1.26	S-D (MA)
24	3.0	.88	3.0	1.03	S-E (MA)
25	3.5	1.20	4.0	.94	G-E (MA)
26	3.6	.79	3.0	1.08	G-D (OC)
27	2.7	.77	2.1	1.02	S-D (OC)
28	3.6	.90	2.7	.76	S-E (OC)
29	2.2	.76	2.2	.81	S-E (OC)
30	2.7	1.00	2.6	.90	S-D (MA)
31	2.1	.88	2.6	.91	G-D (OC)
32	2.8	.93	2.1	.84	S-D (OC)
33	3.8	1.00	2.4	.90	S-E (OC)
34	2.7	.96	3.4	1.00	S-D (MA)
35	3.6	.99	3.7	.92	S-E (MA)
36	3.9	1.03	3.7	1.20	G-D (MA)
37	3.6	.89	3.0	1.08	G-E (MA)
38	3.6	1.10	3.6	1.04	G-E (MA)
39	3.9	.96	4.0	.95	S-E (MA)
40	3.8	.97	2.3	.86	S-D (MA)

Table A6

Summary ANOVA Table

(SRI-A Pilot Rating Form)

Source	SS	df	MS	F	p less than
Content (C)	1.71919	1	1.71919	7.778	.005*
Judgment (J)	.06160	1	.06160	.278	.597
Behavior (B)	1.29398	1	1.29398	5.854	.016*
C x J	.02570	1	.02570	.116	.733
C x B	.45003	1	.45003	2.036	.154
J x B	.16593	1	.16953	.767	.381
C x J x B	.53133	1	.53133	2.404	.121
Residual	84.87072	384	.22102		
Total	89.12208	391	4.47238		

Finally, none of the two-way interactions and the three-way interaction were significant.

Table A7 shows the means and standard deviations of interrater agreement for the four types of items in each content area for the SRI-A pilot rating form. The means were subtracted from 1.0 to reflect the nature of the variable. Accordingly, the higher the mean, the higher the interrater agreement. According to the ANOVA results, items related to the content area 'maintaining attention' are rated with less agreement than items related to 'open communication.' Moreover, items containing general instructor behaviors are also rated with less agreement than items containing specific instructor behaviors.

Development of the Final SRI-A and SRI-B rating forms

The number of items on the pilot rating forms were reduced because most student rating of instruction forms contain about 20 to 30 items. Furthermore, the feedback received during the piloting of the SRI forms indicated that there were too many questions in the same content area and many of these items appeared the same. Therefore, three out of five items in each of the eight subtests were chosen so that the final SRI-A and SRI-B forms consisted of only 24 items each.

To reduce the number of items, each item was reviewed in relation to the other items in the same subtest by the investigator along with an L&ES faculty member who has much experience with student rating of instruction forms. The criteria used to evaluate the items were subjective and objective. The subjective criteria consisted of the appropriateness of the item (content) for a large lecture classroom, its redundancy with other items and the clarity of wording. The

Table A7

Means and Standard Deviations: Interrater Agreement

(SRI-A Pilot Rating Form)

			Open Communication: Content Area		
			Level of Judgment		
			Descriptive	Evaluative	average
LEVEL OF BEHAVIOR	Specific	M	.7805	.6564	.72
		(SD)	(.2983)	(.4340)	
	General	M	.6181	.7285	.67
		(SD)	(.5280)	(.3145)	
			.70	.69	

		Maintaining Attention: Content Area		
		Level of Judgment		
		Descriptive	Evaluative	average
Specific		.6584	.6491	.66
		(.4389)	(.3675)	
General		.5077	.4344	.48
		(.6072)	(.6455)	

objective criteria included the mean and standard deviation of each item. Paired items (i.e., in forms SRI-A and SRI-B) were considered together. The items were evaluated first subjectively and then objectively.

Most items were eliminated on a subjective basis only. When an item on the SRI-A form was eliminated, its parallel item in the SRI-B form was also eliminated.

The remaining items were revised as needed (e.g., all evaluative items were at the "above average" end of the scale). The final items in each subtest for SRI-A and SRI-B are shown in Table A8. An asterisk by the item number (in the final form) indicates that only four out of six judges agreed to its categorization. The final SRI-A and SRI-B rating forms are found at the end of this appendix.

Table A8

Items in Final SRI-A and SRI-B

Rating Forms--by Category

Specific-Descriptive ('Maintaining Attention')

SRI-A

12.* Varied tone of voice.

20. Stated why the subject matter was important.

24. Used variety of instructional activities, media or formats (e.g., guest lectures, panel discussions, etc.).

SRI-B

12. Varied speed of voice.

20. Told students how they could apply the material presented in class to their daily lives.

24. Used audiovisual aids (slides, films, tapes) to illustrate concepts or principles.

Specific-Evaluative ('Maintaining Attention')

SRI-A

4. Was above average in stating what material is to be covered in a lecture.

13.* Was very good at presenting facts and concepts from related fields.

21.* Was above average in using a number of different ways to present a point.

SRI-B

4. Was above average in presenting an overview of a lecture.

13. Was above average at stating how the course material related to your field of interest.

21.* Was very good in using a variety of ways to present the course material.

Table A8 Continued.

General-Descriptive ('Maintaining Attention')

SRI-A

1. Knew when students were bored or confused.

8.* Related subject matter to your interests or activities.

11.* Related the subject matter to other academic disciplines and real world situations.

SRI-B

1. Was aware of students paying or not paying attention in class.

8.* Related class topics to your experiences.

11.* Related course material to real-life situations.

General-Evaluative ('Maintaining Attention')

SRI-A

14. Was very good at maintaining your attention.

22. Was above average in making the course content meaningful.

23. Was above average in enthusiasm about the course.

SRI-B

14. Was very good at motivating the class.

22. Was above average in making the subject matter relevant.

23.* Was above average in being enthusiastic when lecturing.

*Only 4 out of 6 judges agreed to this categorization.

STUDENT REACTIONS TO INSTRUCTION

Form SRI-A

DO NOT WRITE ON THIS FORM

USE SEPARATE ANSWER SHEET

This form allows college students to assist their instructors in improving their teaching. Please give thoughtful and honest responses to the statements which follow. Try to consider each statement separately, rather than let your overall feelings about the instructor determine all the responses. If none of the alternatives appear to express your reaction exactly, respond with the closest appropriate alternative.

Record all responses on the separate answer sheet provided by blackening only the appropriate space with a No. 2 pencil. Be sure to erase errors and stray marks completely.

Read each statement carefully and then indicate the extent to which you agree or disagree with the statement. Use the following code:

1 = if you strongly agree with the statement
2 = if you agree with the statement
3 = if you neither agree nor disagree
4 = if you disagree with the statement
5 = if you strongly disagree with the statement

The Instructor:

1. Knew when students were bored or confused.

2. Responded to the specific question that was asked when answering students' questions.

3. Was above average in encouraging students to participate in class.

4. Was above average in stating what material is to be covered in a lecture.

5. Stated the objectives of each lecture.

6. Answered students' questions very well.

7. Was above average in willingness to explore a variety of points of view.

8. Related subject matter to your interests or activities.

9. Was above average in openness during class discussion.

1 = if you strongly agree with the statement
2 = if you agree with the statement
3 = if you neither agree nor disagree
4 = if you disagree with the statement
5 = if you strongly disagree with the statement

10. Encouraged students to participate in class discussion.

11. Related the subject matter to other academic disciplines and real world situations.

12. Varied tone of voice.

13. Was very good at presenting facts and concepts from related fields.

14. Was very good at maintaining your attention.

15. Expected students to answer questions in class.

16. Told students which topics were most important and which were least important.

17. Was very good at asking questions of students.

18. Stated very well which topics were important.

19. Tolerated other points of view.

20. Stated why the subject matter was important.

21. Was above average in using a number of different ways to present a point.

22. Was above average in making the course content meaningful.

23. Was above average in enthusiasm about the course.

24. Used a variety of instructional activities, media or formats (e.g., guest lectures, panel discussions, etc.).

STUDENT REACTIONS TO INSTRUCTION

Form SRI-B

DO NOT WRITE ON THIS FORM

USE SEPARATE ANSWER SHEET

This form allows college students to assist their instructors in improving their teaching. Please give thoughtful and honest responses to the statements which follow. Try to consider each statement separately, rather than let your overall feelings about the instructor determine all the responses. If none of the alternatives appear to express your reaction exactly, respond with the closest appropriate alternative.

Record all responses on a separate answer sheet provided by blackening only the appropriate space with a No. 2 pencil. Be sure to erase errors and stray marks completely.

Read each statement carefully and then indicate the extent to which you agree or disagree with the statement. Use the following code:

1 = if you strongly agree with the statement
2 = if you agree with the statement
3 = if you neither agree nor disagree
4 = if you disagree with the statement
5 = if you strongly disagree with the statement

The Instructor:

1. Was aware of students paying or not paying attention in class.

2. Asked students if they had any questions on the material read or about previous lecture.

3. Was extremely good in encouraging student participation.

4. Was above average in presenting an overview of a lecture.

5. Explained the course objectives.

6. Was very good at answering questions from students.

7. Was very good in being open to students' viewpoints.

8. Related class topics to your experiences.

9. Was above average in being open when answering questions from students.

1 = if you strongly agree with the statement
2 = if you agree with the statement
3 = if you neither agree nor disagree
4 = if you disagree with the statement
5 = if you strongly disagree with the statement

10. Promoted teacher-student discussion (as opposed to mere responses to questions).

11. Related course material to real-life situations.

12. Varied speed of voice.

13. Was above average at stating how the course material related to your field of interest.

14. Was very good at motivating the class.

15. Encouraged silent students to participate.

16. Stated what was important to learn in each class session.

17. Was above average in questioning the class.

18. Stated the objectives of the course very well.

19. Was willing to explore a variety of points of view.

20. Told students how they could apply the material presented in class to their daily lives.

21. Was very good in using a variety of ways to present the course material.

22. Was above average in making the subject matter relevant.

23. Was above average in being enthusiastic when lecturing.

24. Used audiovisual aids (slides, films, tapes) to illustrate concepts or principles.