12-1-1999

Rater Training to Improve Student Evaluations of Teaching Effectiveness

Robert Beeler
Western Kentucky University

Follow this and additional works at: http://digitalcommons.wku.edu/theses
Part of the Education Commons, and the Psychology Commons

Recommended Citation
http://digitalcommons.wku.edu/theses/743

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact topscholar@wku.edu.
RATER TRAINING TO IMPROVE STUDENT EVALUATIONS
OF TEACHING EFFECTIVENESS

A Thesis

Presented to

the Faculty of the Department of Psychology

Western Kentucky University

Bowling Green, Kentucky

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

Robert Neal Beeler

December 1999
RATER TRAINING TO IMPROVE STUDENT EVALUATIONS
OF TEACHING EFFECTIVENESS

Date Recommended  7-08-99

Elizabeth L. Shouflet
Director of Thesis

John O'Connor

Elmer Gray  12/07/99
Dean, Graduate Studies and Research  Date
Acknowledgments

I would like to thank Dr. Betsy Shoenfelt, my advisor on this project, for her continued support and motivation. I would also like to thank Drs. James Craig and John O’Connor for agreeing to advise this thesis. Their input and direction helped to keep this project on track. Finally, this project would not have been possible without the support of my parents, fiancee and friends who have motivated me to “keep going.” The light at the end of the tunnel has been their shining love and support.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>iii</td>
</tr>
<tr>
<td>Figure Titles</td>
<td>v</td>
</tr>
<tr>
<td>Abstract</td>
<td>vi</td>
</tr>
<tr>
<td>Chapter 1: Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Purposes of Student Evaluations</td>
<td>2</td>
</tr>
<tr>
<td>Validity of Student Evaluations</td>
<td>6</td>
</tr>
<tr>
<td>Measures of Teaching Effectiveness</td>
<td>7</td>
</tr>
<tr>
<td>Criterion-related Validity</td>
<td>10</td>
</tr>
<tr>
<td>Content Validity</td>
<td>13</td>
</tr>
<tr>
<td>Chapter 2: Rater Training</td>
<td>15</td>
</tr>
<tr>
<td>Chapter 3: Video Tape as a Training Tool</td>
<td>18</td>
</tr>
<tr>
<td>Chapter 4: The Present Study</td>
<td>20</td>
</tr>
<tr>
<td>Method</td>
<td>21</td>
</tr>
<tr>
<td>Results</td>
<td>25</td>
</tr>
<tr>
<td>Discussion</td>
<td>33</td>
</tr>
<tr>
<td>References</td>
<td>39</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Table 1: Dimensions of Teaching Performance</td>
<td>12</td>
</tr>
<tr>
<td>2. Table 2: Correlations of scores taken before and after intervention</td>
<td>26</td>
</tr>
<tr>
<td>3. Table 3: Correlations of scores for Trained Groups</td>
<td>27</td>
</tr>
<tr>
<td>4. Table 4: Correlations of scores for Control Group</td>
<td>27</td>
</tr>
<tr>
<td>5. Table 5: Correlations between measures of student grade</td>
<td>28</td>
</tr>
<tr>
<td>6. Table 6: Dimension Means and Standard Deviations</td>
<td>30</td>
</tr>
<tr>
<td>7. Table 7: Dimension Correlations with Student Grade</td>
<td>31</td>
</tr>
<tr>
<td>8. Table 8: One way ANOVA of difference scores across groups</td>
<td>32</td>
</tr>
<tr>
<td>9. Table 9: Item variance accounted for by Global Measure and Grade</td>
<td>33</td>
</tr>
</tbody>
</table>
RATER TRAINING TO IMPROVE STUDENT EVALUATIONS
OF TEACHING EFFECTIVENESS

Robert Neal Beeler December 1999 57 Pages

Directed by: Elizabeth Shoenfelt, Ph.D., James Craig, Ph.D., and John O’Connor, Ph.D.

Department of Psychology Western Kentucky University

The impact of rater training on students performing evaluations of teaching
effectiveness was evaluated. Training was administered either via live lecture or video
tape. Both trained groups were compared with a no training control group. It was
hypothesized that the training would result in more accurate assessments of teaching
evaluations and that there would be no significant differences between either of the trained
groups. Results showed that students were able to distinguish between dimensions of
performance. However, no effects of training were found. The trained groups did not
differ significantly from the untrained control group on instructor ratings.
Chapter 1

Introduction

In this study I examined the impact of training on increasing the accuracy of student ratings of teaching effectiveness. Additionally, the differences between videotaped training versus lecture-based training were explored. It was hypothesized that training student raters to more accurately assess teaching effectiveness would increase the utility of student evaluations of teaching. Currently, student ratings of teaching behavior play a role in institutional judgement of instructional performance; the size of this role is dictated by the institution. However, student measures may not be a valid indication of classroom behavior if such measures do not accurately measure teaching effectiveness.

The use of students' ratings as an indication of faculty effectiveness is accepted, at least in part, at many educational institutions. Students have the greatest exposure to faculty in terms of instructional time and thus may provide the institution with a useful evaluative benchmark regarding the effectiveness of instruction. However, the ratings provided by students are not always an accurate depiction of instructional effectiveness. Student ratings of professors may represent only the popularity or the enthusiastic presentation style of the instructor (Abrami, 1989). If the accuracy and utility of such measures could be improved, then a more accurate, comprehensive evaluation of faculty
can be made. Rater training has been identified by Smith (1986) and Woehr (1994) as a tool for improving the effectiveness of performance ratings.

To increase the accuracy of student ratings of instructional effectiveness, students were trained on the various dimensions of teaching performance and on common errors made in rating performance (Smith, 1986; Woehr, 1994). Smith (1986) stated that rater training may improve the effectiveness of performance ratings. Specifically, frame of reference training when combined with performance dimension training is effective at reducing leniency and other rater errors. Smith (1986) concluded "the best way to increase rating accuracy is to combine dimension training with frame of reference training (p. 35)." Dimention training identifies the dimensions of behavior that comprise effective job performance. Rater error training, an effective tool for reducing the number of errors made in rating performance (Woehr, 1994), defines common errors made in performance ratings. Training students to accurately rate teaching performance should increase the utility of student evaluations of teaching. Improving teaching evaluations should be helpful to the instructor and the institution, as well as to the student.

**Purposes of Student Evaluations**

Numerous educational institutions utilize student ratings as a measure of overall faculty effectiveness. In fact, at many such institutions student ratings are the only source of information regarding instructor classroom performance. Marsh (1987) identified five reasons for collecting student ratings of faculty effectiveness. These are identified below and each will be discussed.
1. diagnostic feedback to faculty about the effectiveness of their teaching to be used for improving teaching (e.g., techniques)

2. a measure of teaching effectiveness used for administrative decision making

3. a measure of student opinion of courses/instructors for students’ use in course selection

4. direction and feedback for course and curriculum development; a measure of course quality

5. direction for further research on faculty and teaching effectiveness

First, student evaluations provide the instructor with feedback necessary for improving teaching performance (Marsh, 1987; Marsh, 1984; Costen, Greenough, & Menges, 1971). Specific student feedback on instructional technique is particularly useful to faculty with less educational tenure. Inexperienced faculty can benefit from feedback differentiating effective and ineffective classroom behavior. Although Borman, Hanson, Oppler, and Pulakos (1993) found experience to have more impact on performance than feedback alone, student ratings may identify instructional skills gaps to all faculty, which in turn may provide the student with a more effective learning experience.

Second, at many institutions student evaluations play some role in the determination of overall faculty effectiveness judgments used in such administrative-based decisions as promotion and tenure. As classroom performance is an area where many faculty spend a majority of their time, either in actual instruction or in preparation, evaluations of teaching rightly play a role in judging faculty performance. Student
evaluations provide the institution with an evaluation of teachers' classroom performance. Student input is important in obtaining a more complete view of the teacher and classroom performance since students have the greatest exposure to instruction. However, the student view of classroom performance may not be unbiased. If students are aware that their ratings are used for administrative or promotional decisions, the ratings are more likely to be inflated (Frase & Streshly, 1994). Such ratings carry little utility. The accuracy of student evaluations must be assured if they are to be useful to the instructor or the institution.

As stated in Marsh's (1987) third purpose, some institutions make student evaluations public for inspection by prospective students in selecting instructors and/or courses. Toward this purpose, student evaluations serve the students rather than the institution by providing information that is useful to the student. Abrami, d'Apollonia, & Cohen (1990) stated student evaluations serving this purpose portray only the popularity or leniency of the instructor and/or course. For example, an instructor who grades strictly may receive less favorable student ratings than an instructor who grades leniently even though the students in the more difficult instructor's class may have learned more in that course. Student evaluations serving this purpose are not an accurate depiction of instructional effectiveness (Abrami, et al., 1990).

Abrami et al. (1990) stated that student achievement is the most accurate measure of instructional effectiveness. Evaluations that focus on individual achievement rather than the student view of the instructor/course effectiveness are a more objective, useful measure of teaching effectiveness (Marsh, 1987). Abrami et al. (1990) agreed that student
evaluations of instructional effectiveness are valid if they address student achievement. Marsh (1987) explained if students are taught the same material by different instructors, the effectiveness of the instructor could be measured by student grade. Effective teaching would result in a higher average student grade and a higher average student rating of teaching effectiveness. Students learn more if they are taught effectively.

However, evaluating instructional effectiveness may not be the goal of student evaluations. Abrami, et al. (1990) found support for the validity of student ratings reflecting the opinions of students in the classroom. If student evaluations serve the purpose of informing future students of the student view of the course/instructor, student ratings would likely be a valid indication of cumulative student opinion. However, the validity of student opinions should not be confused with the validity of student evaluations of teaching, which serve the purpose of measuring teaching effectiveness.

Marsh’s fourth purpose for conducting student evaluations of teachers is to further develop the course or curriculum of the department or institution. Student evaluations can identify the areas of curriculum that are lacking. Proper changes can then be made to further develop the course and curriculum of the institution to better educate the students. Because students have the greatest exposure to instructional time, they may be able to provide a useful tool to evaluate the course or the curriculum. However, because student evaluations are more a function of the instructor and not of course content, student evaluations are more useful to the instructor and less useful for providing information relevant to the curriculum (Marsh, 1987). Little support has been found for student
evaluations collected for the purpose of course and/or curriculum development (Marsh, 1987).

The final purpose for conducting student evaluations, as stated by Marsh, is to further research on faculty and teaching effectiveness. If student evaluations can accurately identify effective instruction, then these evaluations can serve as criterion measures in determining factors to improve the effectiveness of instruction. Further research may be conducted to ascertain any moderating variables impacting instructional effectiveness. The identified effective modes of instruction may be shared with others to increase the effectiveness of teaching.

Most of Marsh’s purposes for conducting student evaluations address the effectiveness of teaching. The third purpose, however, may address student opinion of the instructor and/or the course and not necessarily the effectiveness of teaching. Because the focus of this study is on teaching effectiveness, the third objective is not relevant. The focus of this study was on only the student evaluations conducted for purposes of measuring teaching effectiveness. It is essential to obtain an accurate measure of effective teaching. Student ratings are perhaps the best medium for obtaining such measures. Student evaluations, at a minimum, play some role in the judgement of overall faculty effectiveness at many educational institutions. However, an accurate view of classroom performance is necessary to fairly judge faculty effectiveness.

Validity of Student Evaluations

Muchinsky (1997) defined validity as the “correctness” (i.e., accuracy and precision) of the measure. Crittenden and Norr (1975) suggested that “a valid measure (of
student evaluations) would be one which correlates with factors theoretically related to teaching effectiveness . . . (p. 431).” Feldman (1976) and Abrami, et al. (1990) agreed that student ratings maintain their validity if they accurately reflect instructional effectiveness. To assess the validity of student evaluations, the evaluations must be correlated with another measure of instructional effectiveness. Convergent validity is found when measures collected through one method are correlated with measures collected through another method.

**Measures of Teaching Effectiveness**

Marsh (1987) believed methods other than student evaluations can be used to measure teaching effectiveness. Specifically, student learning and instructor self-evaluations are recognized methods of assessing teaching effectiveness (e.g., Marsh, 1987; Schneider, Hanges, Goldstein, & Braverman, 1994; Crittenden & Norr, 1975). Student evaluation scores can be correlated with course grade and instructor self-evaluation scores to accurately measure teaching effectiveness.

**Student Achievement.** The most widely accepted measure of effective teaching is student learning (Day, 1995; Frey, Leonard, & Beatty, 1975; Feldman, 1976; Marsh, 1987; Warrington, 1973); that is, effective instruction should result in higher mean student learning (i.e., grades) than should less effective instruction. According to this line of reasoning, student ratings of teaching effectiveness are valid if there is a positive correlation with student grade.

However, Marsh (1987) cautioned that student learning is not synonymous with effective teaching. Higher grades in a course may not be the product of effective teaching.
Grades may be attributed to something other than effective instruction, such as student motivation or leniency of the instructor. Feldman (1976) and McKeachie (1997) stated that grades may have an impact on the outcome of student evaluations. Feldman (1976) found a positive correlation between expected grades and instructor evaluation. Feldman did not find a causal relationship between grades and evaluation scores. Feldman believed it possible that students who received good grades could reward the instructor with a good evaluation. Likewise, McKeachie (1997) found that "giving higher grades can raise ratings (p. 1220.)." Despite these criticisms, Frey et. al (1975) believed that student accomplishment, controlling for student aptitude (i.e., scores on an objective exam administered prior to the semester), is the best predictor of student ratings. By controlling for student aptitude, Frey et. al were able to obtain a more accurate measure of performance gain. Likewise, Crittenden & Norr (1975) found student performance gain to be the most "theoretically desirable" criterion against which to validate student ratings of teaching effectiveness (p. 431).

Instructor Self-Evaluations. Marsh (1987) found significant convergent validity coefficients between student ratings and instructor self-evaluations on each dimension of teaching (r=0.49, corrected and 0.45, uncorrected). Furthermore, the correlations between student ratings and instructor self-evaluations were higher on the same dimension than they were on different dimensions, demonstrating the ability of the instructor (and student) to differentiate between dimensions of performance. Schneider et al. (1994) also found significant correlations between student evaluations and instructor self-evaluations, illustrating the adequacy of self-evaluations in judging the accuracy (i.e., validity) of
student ratings. Marsh explained that the agreement between student ratings and instructor self-evaluations "demonstrates the validity" of self-evaluations in predicting student ratings; thus, one method of addressing the accuracy of student ratings would be to assess their relationship to instructor self-evaluations.

**Student Evaluations.** In the simplest form, effective teaching is measured on a global scale. For example, "Overall, my teacher is an effective teacher." Such an item would encompass any and all dimensions of teaching performance. Using only one item to assess teaching performance limits the rater to only a global, overall impression of behavior. No specific feedback is given to the instructor, limiting the utility of the measure. Because there is no differentiation among dimensions of teaching, faculty cannot gain specific, dimensional performance information. Yet, Abrami (1989) concluded that a single measure of effectiveness is the most useful and most accurate portrayal of performance. Day (1995) found an overall impression rating of teaching performance to account for a significant amount of variance when evaluating classroom performance. Proponents of the global measure of performance believe raters are unable to differentiate significantly between the different dimensions of performance (Abrami, 1989; Abrami et al. 1990). However, Day (1995) found that students who were made aware of the different dimensions of performance rated instructors more accurately than did students without dimension knowledge.

Effective performance cannot always be reflected in a single item. Because all dimensions of performance may not be addressed, some (e.g., Marsh, 1987; Frey et al., 1975; Cashin & Downey, 1992) believe that a single measure does not allow for faculty
comparisons. In addition, general performance feedback is not specific enough to aid professors in correcting specific, ineffective performance. Measuring performance on several dimensions allows for specific feedback to instructors regarding classroom behavior, which in turn facilitates classroom performance improvement efforts. For example, instructors rated low in the area of grading can spend their efforts improving that area of teaching rather than devoting effort to a dimension of performance in which they excel. Rushton, Brainerd, and Pressley (1983) concluded that an overall measure of performance is less reliable and less valid than a measure of performance on more than one dimension. Marsh (1984) stated that there is no single dimension of effective teaching. Rather, effective teaching is the result of performance on several dimensions, which when combined make up overall teaching performance. Teachers effective on one dimension of teaching (e.g., organization) are not necessarily effective on another (e.g., enthusiasm) (Marsh & Roche, 1997). In order to evaluate the effectiveness of teaching, all facets of teaching performance must be accounted for. Many researchers (Day, 1995; Cashin & Downey, 1992; Feldman, 1997; Marsh & Roche, 1997) subscribe to the principle of multiple dimensions of performance. Researchers subscribing to this theory believe effective instruction to be the result of several dimension behaviors. Effective instruction can determined by several methods. A criterion-related validity approach was used to identify the criteria that will be used to evaluate teaching performance in the present study.

**Criterion-related validity**

Relevant criteria that capture the dimensions of effective teaching must be identified. A criterion-related validity approach identifies a standard or criterion of
performance that is correlated with another measure of performance effectiveness (Muchinsky, 1997).

Much research has concentrated on the multidimensionality of performance (Feldman, 1976; Frey et al., 1975; Hildebrand, Wilson, & Dienst, 1971; Marsh, 1982; Warrington, 1973). Researchers vary in their conclusions of how many dimensions are needed to effectively capture teaching performance. Estimates range from five to nineteen dimensions of performance (Feldman, 1976; Frey et al., 1975; Hildebrand et al., 1971; Marsh, 1982, Warrington, 1973). Frey et al. classified instruction into seven independent dimensions (see Table 1). Frey et al. believed that to assess effective classroom performance the evaluation instrument should address each of these dimensions. The seven dimensions identified by Frey et al. are consistent with most of the dimension taxonomies identified by other researchers (see Table 1).

Hildebrand et al., (1971) and Warrington (1973) both identified five dimensions of teaching, only three of which were in common (see Table 1). The dimensions identified by Hildebrand et al. and Warrington, as well as the seven dimensions identified by Frey et. al (1975), are included in the dimensions identified by Feldman (1976). Feldman identified 19 dimensions which differentiated superior teaching from teaching considered to be less than superior. Marsh (1987) considered Feldman's 19 dimensions to be the most comprehensive list of teaching dimensions. However, Marsh cautioned that students may not be capable of distinguishing between 19 dimensions.
Table 1
Researchers Dimensions of Effective Teaching Performance

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization / Planning</td>
<td>Preparation / Organization</td>
<td>Organization / Clarity</td>
<td>Organization</td>
<td>Organization</td>
</tr>
<tr>
<td>Presentation Clarity</td>
<td>Clarity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workload</td>
<td>Difficulty / Workload</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal Attention</td>
<td>Respect for Students</td>
<td>Individual Interaction</td>
<td>Individual Interaction</td>
<td>Individual Rapport</td>
</tr>
<tr>
<td>Classroom Discussion</td>
<td>Discussion</td>
<td>Interaction</td>
<td>Instructor Involvement</td>
<td>Group Interaction</td>
</tr>
<tr>
<td>Course Material</td>
<td></td>
<td>Assignments / Readings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grading</td>
<td>Exams</td>
<td></td>
<td></td>
<td>Exams / Grading</td>
</tr>
<tr>
<td>Student Accomplishment</td>
<td>Sensitive to Progress</td>
<td>Student Interest / Performance</td>
<td></td>
<td>Learning</td>
</tr>
<tr>
<td>Personal Interest</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enthusiasm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breadth</td>
<td>Analytic / Synthesis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teaching Aids</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feedback</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Challenge</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Availability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Each column represents one evaluative tool from a researcher or group of researchers. Each row represents a dimension of teaching performance identified by the evaluative instrument. Shaded rows depict the nine dimensions shared by Marsh (1987) and at least two other qualified instruments.
To simplify, Marsh (1987) consolidated the work of several researchers (Frey et al., 1975; Feldman, 1976; Hildebrand, et al., 1971; Warrington, 1973) into nine dimensions. Marsh’s nine dimensions were identified through an analysis of current evaluation instruments, student and teacher interviews, and through a psychometric analysis of evaluation instrument properties. Marsh found factor analytic support for the multidimensionality of the instruments identified in Table 1.

Marsh’s nine dimensions of performance (1982, 1987) seem to be the most agreed upon dimensions for measuring faculty effectiveness. Each of these nine dimensions is included by at least two other validated faculty evaluation instruments (Frey et al., 1975; Feldman, 1976; Hildebrand et al., 1971; Warrington, 1973; Marsh, 1987). An additional item measured effective teaching performance on a global scale. Table 1 identifies the number of dimensions supported by various researchers and how they relate to one another.

Content validity

According to the Encyclopedia of Educational Evaluation (1975), content validity “refers to the degree to which a measure captures a program objective (p. 460).” Content validity places relevant criteria into the context of an instrument used to evaluate the concept or idea of effective teaching. To accurately measure effective teaching, all relevant dimensions of effective teaching must be represented on the instrument. The validity and usefulness of effective measures of teaching are contingent on the content and coverage of the specific items on the instrument. Thus, content validity is the extent to which the evaluation instrument items measure effective teaching. A valid teaching
evaluation instrument should encompass each of the nine dimensions identified by Marsh (1987).

In sum, the validity of measures of instructor effectiveness can be evaluated by correlating student ratings of performance with student grades and instructor self-evaluations. To ensure content validity, evaluation items measuring classroom performance should address the nine dimensions of performance identified by Marsh (1987). Likewise, due to the potential impact of grades on student ratings (Crittenden & Nott, 1975; Feldman, 1976; McKeachie, 1997), student achievement in the course should be controlled.
Chapter 2

Rater Training

Rater training has been identified by Smith (1986) and Woehr (1994) as a tool for improving the effectiveness of performance ratings. The majority of performance evaluation research identifies rater training as the most cost effective method of reducing rating errors and increasing rater accuracy (Lee, 1994; Woehr, 1994; Woehr & Huffcutt, 1994; Pelley, 1991; Brecker, 1989; Hedge & Kavanaugh, 1988; Smith, 1986; Lee, 1985; McIntyre, Smith, & Hassett, 1984). Costin, Greenough, & Menges (1971) found training to be effective at increasing the accuracy of instructional effectiveness measures. Four types of rater training programs have been identified: error training, frame-of-reference training, performance dimension training, and behavior observation training.

Rater Error Training

Rater error training concentrates on identifying and reducing common rating errors such as leniency or severity errors. Leniency or severity errors are present when raters tend to rate performers higher or lower than the level of performance demonstrated by their behavior, resulting in inaccurate ratings. Rater error training programs have been empirically proven to reduce errors made in rating performance. However, while rater error training is an effective method of reducing the number of rating errors, it has little to
no impact on the accuracy of performance measures (Hedge & Kavanaugh, 1988; Roth, 1988; Smith 1986; & Woehr & Huffcutt, 1994; Zedeck & Cascio, 1982).

**Performance Dimension Training**

A second type of rater training program is performance dimension training. The goal of performance dimension training is to identify the various dimensions that constitute performance. Halo error is the generalization of evaluations (good or bad) based on a global impression or on one dimension of performance to all dimensions of performance. Halo errors can be prevented with dimension training (Smith, 1986). If raters are aware of the various dimensions of performance, then independent judgements can be made on each dimension, giving a more accurate depiction of overall performance. Performance dimension training is quite successful in improving the accuracy of ratings (Woehr & Huffcutt, 1994; Smith, 1986).

**Frame-of-Reference Training**

Frame-of-reference (FOR) training provides the raters with a standard by which performance may be judged. If raters are given an example of good and poor performance in each dimension, they may then judge more accurately the performance of ratees. Woehr (1994; Woehr & Huffcutt, 1994) stated that FOR training is the single most effective training strategy for increasing rater accuracy. Smith (1986) and Day & Sulsky (1995) concurred that FOR training is an effective method for increasing rater accuracy.

**Behavior Observation Training**

The final training strategy is behavior observation training. Behavior observation training is a training strategy whereby raters are given cues or specific examples of
behaviors that differentiate effective and ineffective performance on the job. Behavior observation training teaches raters how to observe performance and how to differentiate good and poor performance. Although not as popular as the other types of rater training strategies, behavior observation training has the potential to be very effective in reducing rating errors as well as increase rating accuracy (Smith, 1986; Woehr & Huffcutt, 1994).

Present Study Training Strategy

Based on the research regarding the effects of training on ratings, training students to more accurately perform teacher evaluations should encompass error training, performance dimension training, and frame-of-reference training. Rater error training should decrease the number of rating errors made by students. However, rater error training has little impact on the accuracy of student ratings (Hedge & Kavanaugh, 1988; Smith 1986; Woehr & Huffcutt, 1994). Marsh (1987) identified performance dimensional training, which stresses the importance of differentiating the dimensions of performance, as a useful tool for improving the accuracy of performance measures. Raters should be made aware of the nine dimensions that constitute effective teaching. Frame-of-reference training will provide students with an example of good and poor performance on each dimension of effective teaching. Frame-of-reference training is, perhaps, the single most effective training strategy for increasing rating accuracy (Woehr, 1994; Woehr & Huffcutt, 1994). Rater training should both decrease the number of rating errors made and increase the accuracy of the rating.
Chapter 3

Video Tape as a Training Tool

The use of video tape as a training tool has grown since its invention in the early 1960s. Video tape is an excellent tool for assuring standardized presentation of material. Training via video tape is a method by which many trainees can be trained at the same time as many subjects can view a single television simultaneously. Through the use of video tape, training programs are more consistent, lower in cost, and more convenient than programs delivered through other media (Bove, 1984; Cartwright, 1986).

Video taped training, however, is not without its limitations (Cartwright, 1986; Chu & Schramm, 1968; Bove, 1984). The initial start-up costs involved with producing a video taped training program are often high. The necessary equipment for developing the video may not be available. If out-sourced, or contracted from another firm, updates are often difficult to achieve, jeopardizing the currency of the training program. Also, it is important to choose a presenter viewed as credible by the audience (Cartwright, 1986). In addition, skilled individuals are need to write the training script, produce the video, and direct the programming (Cartwright, 1986).

Despite its disadvantages, the use of video-based training is very effective. Video training was found to be at least as effective as lecture training (Chu & Schramm, 1968).
When there is no discussion or practice by trainees in lecture-delivered or video taped-delivered groups, Schramm (1972) found video tapes to be as effective a medium for delivering training as lecture-based training. However, Dwyer (1972) found the effectiveness of any training program was increased by asking probing questions before the training session. Furthermore, Schramm (1972) found that the use of humor or elaborate visuals do not increase the effectiveness of video-based training. Chu and Schramm (1968) concluded that varying the style of presentation also was ineffective at increasing the effectiveness of video-based training.
Chapter 4

The Present Study

The present researcher evaluated the effectiveness of rater training for improving the accuracy of student evaluations of faculty teaching. Training was provided through live lecture and video tape. Both groups were compared with a no training control group.

Teaching evaluations were made using a rating form adopted from Purdue University in 1977. The form includes six core items representing five dimensions of teaching. Four additional items were added to encompass the nine dimensions that constitute effective teaching (Marsh, 1987). Each dimension was represented by a single item on the university teaching evaluation form. An additional item measuring the overall performance of the instructor was also included.

Students were trained to accurately evaluate their instructors. The training program defined common rating errors. The rater training also included a definition and frame of reference for each dimension of effective teaching identified by Marsh (1987). A script of the training program can be found in Appendix A. The same script was used for the lecture and video taped training groups. Two trained groups were compared with a third, no training, control group. The two trained groups received the same training via two different mediums. A lecture-trained group was compared to a group trained via video tape to assess the effectiveness of video tape as a substitute for lecture training.
Hypotheses.

In the present study, the following relationships were examined.

Hypothesis 1: Student evaluation scores of the trained groups should be more closely related to student grade for each dimension of teaching than should the no training group.

Hypothesis 2: Student evaluation scores of the trained groups should be more closely related to instructor self-evaluation for each dimension of teaching than should the no training group.

Hypothesis 3: There will be no difference between the two trained groups on the relationship between student evaluations and student grade for each dimension of performance.

Hypothesis 4: There will be no difference between the two trained groups on the relationship between student evaluations and instructor self-evaluation for each dimension of performance.

Method

Participants

One hundred and thirty-four students from six introductory psychology classes from a large, southeastern state university participated in the study. Three instructors each taught two sections of the course. Ninety-three percent of the participants were White (n=125), six percent were Black (n=8), and one percent were Asian (n=1). Fifty-nine percent of the participants were female (n=79). The mean age of participants was 19.2 years (SD=1.19).
Materials

Rater Training. A training script was developed for discussing three training types: performance dimension training, frame-of-reference training, and rater error training. The training session, developed by a member of the graduate faculty, discussed common rater errors and provided definitions and examples for each dimension of effective teaching performance.

The majority of the training time was focused on dimension training and providing the students with a frame-of-reference for each dimension of performance. Each dimension of performance was defined, followed by an example of good and poor performance. Marsh (1987) discussed the importance of differentiating the dimensions of performance to students to improve the accuracy of performance measures. The nine dimensions of effective classroom performance are organization, workload/difficulty, individual rapport, group interaction, course demands, exams/grading, learning, enthusiasm, and breadth (Marsh, 1987).

Additionally, rater error training was included in the training session. Woehr (1994) identified rater error training as an effective method of reducing the number of errors made when rating performance. The training session defined such common rating errors as leniency, severity, and halo and provided an example of each type of error.

After the training script was developed, the university’s communications department made a video tape of the session. The video tape was simply a recording of the lecture training session. A complete script of the training program can be found in Appendix A.
**Rating Instrument.** The rating scale used was an appended version of the form used by the university where the study was conducted. Nine items respectively measured the nine dimensions of performance identified by Marsh (1987). An additional item measured the overall effectiveness of the instructor. The 10 items are listed below. The dimension of performance addressed by each item follows in parentheses.

1. My instructor displays a clear understanding of course topics. (Breadth)
2. My instructor displays interest in teaching this class. (Enthusiasm)
3. My instructor is well-prepared for class. (Organization)
4. Performance measures (exams, assignments, etc.) are well-constructed. (Exams)
5. My instructor is actively helpful. (Individual Interaction)
6. Relative to other courses, the course workload was light. Workload/Difficulty)
7. Students were encouraged to ask questions and were given meaningful answers. (Group Discussion/Interaction)
8. Readings, homework, etc. contributed to the appreciation and understanding of the subject. (Course Demands/Materials)
9. I have learned and understood the subject materials in this course. (Student Performance/Learning)
10. Overall, my instructor is effective. (Global Measure)

Instructors evaluated their own performance using the same form used by the students.

**Procedure**

Each semester students at the university where this study was conducted are asked to evaluate their professors. The university uses these ratings to judge faculty for administrative decisions. Additionally, the ratings are made available to the instructor for individual development purposes. Ratings collected for the purpose of this study were to be used for research purposes only. The university collected additional ratings for the standard evaluations.
teaches more than one course. Despite consistent material and instructional techniques, instructional effectiveness is judged on the basis of the audience, which differs across classes. What is judged to be effective by one class may not be judged as effective in another. Therefore, students in each class were randomly divided into one of three groups. All ratings were performed inside the regular time limits of class. Evaluations were conducted approximately 12 weeks into the 16 week long semester.

Students were identified as “A,” “B,” “C,” “D,” or “below” students based on grades from tests and homework scores completed during the semester. Students with a cumulative grade between 90-100 were classified as a one. Students whose grade fell between 80-89 were classified as a two, between 70-79 a three, 60-69 a four, and below 60 a five. To control for the potential impact of expected grades on ratings (McKeachie, 1997), class members were randomly assigned to one of three treatment groups controlling for grades (i.e., an equal number of “A” students, “B” students, etc. were assigned to each of the three groups).

Following the assignment of all students to one of three experimental groups, each group was taken to a separate room so other students would not overhear any of the training information given to other groups. After groups were separated in each of their rooms, an initial rating of the instructor was obtained to serve as the “pre-training” measure. Following the initial evaluation, group one received live lecture training covering rater errors, dimensions of performance, and frame-of-reference. Group two viewed a video tape of the same training session. Following the training session, students were asked to again evaluate their instructor’s classroom performance. Standardized
instructions were given to all students to limit potential confounding. A script of all the instructions can be found in Appendix B. Because the control group, group three, received no training they were instructed to re-evaluate their instructor’s performance immediately following the initial evaluation. The second evaluation instrument included the same ten performance rating items as on the first instrument presented in a different order. A copy of both evaluation instruments can be found in Appendix C.

Results

Comparison of Pre-/Post-Intervention Scores

Initially, pre-training evaluation scores were correlated with post-intervention evaluation scores across all groups. A high intra-dimensional correlation would illustrate the ability of the students to differentiate between dimensions. Table 2 shows the correlations between evaluation scores taken before and after the intervention. In each case, the strongest correlation was found within the same dimension before and after the intervention. A strong intra-dimension correlation illustrates the ability of the students to differentiate between the dimensions of teaching performance. Because one of the objectives of the training was to teach students to discriminate between dimensions of performance, the same correlation matrix was calculated separately for the two trained groups and the control group. The results can be seen in Tables 3 and 4.
<table>
<thead>
<tr>
<th></th>
<th>Pre1</th>
<th>Pre2</th>
<th>Pre3</th>
<th>Pre4</th>
<th>Pre5</th>
<th>Pre6</th>
<th>Pre7</th>
<th>Pre8</th>
<th>Pre9</th>
<th>Pre10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post1</td>
<td>.485*</td>
<td>.406*</td>
<td>.441*</td>
<td>.416*</td>
<td>.444*</td>
<td>.027</td>
<td>.416*</td>
<td>.238*</td>
<td>.451*</td>
<td>.494*</td>
</tr>
<tr>
<td>Post2</td>
<td>.430*</td>
<td>.736*</td>
<td>.357*</td>
<td>.283*</td>
<td>.438*</td>
<td>.096</td>
<td>.392*</td>
<td>.250*</td>
<td>.369*</td>
<td>.436*</td>
</tr>
<tr>
<td>Post3</td>
<td>.327*</td>
<td>.189</td>
<td>.768*</td>
<td>.440*</td>
<td>.263*</td>
<td>.222</td>
<td>.167</td>
<td>.081</td>
<td>.329*</td>
<td>.447*</td>
</tr>
<tr>
<td>Post4</td>
<td>.280*</td>
<td>.146</td>
<td>.453*</td>
<td>.702*</td>
<td>.349*</td>
<td>.171</td>
<td>.202</td>
<td>.144</td>
<td>.548*</td>
<td>.502*</td>
</tr>
<tr>
<td>Post5</td>
<td>.299*</td>
<td>.436*</td>
<td>.262*</td>
<td>.403*</td>
<td>.705*</td>
<td>-.132</td>
<td>.393*</td>
<td>.330*</td>
<td>.474*</td>
<td>.463*</td>
</tr>
<tr>
<td>Post6</td>
<td>-.033</td>
<td>-.025</td>
<td>.301*</td>
<td>.134</td>
<td>-.170</td>
<td>.918*</td>
<td>-.070</td>
<td>-.034</td>
<td>.067</td>
<td>.051</td>
</tr>
<tr>
<td>Post7</td>
<td>.298*</td>
<td>.382*</td>
<td>.249*</td>
<td>.248*</td>
<td>.500*</td>
<td>-.052</td>
<td>.680*</td>
<td>.309*</td>
<td>.308*</td>
<td>.407*</td>
</tr>
<tr>
<td>Post8</td>
<td>.208</td>
<td>.252*</td>
<td>.159</td>
<td>.353*</td>
<td>.522*</td>
<td>-.035</td>
<td>.327*</td>
<td>.697*</td>
<td>.361*</td>
<td>.448*</td>
</tr>
<tr>
<td>Post9</td>
<td>.335*</td>
<td>.270*</td>
<td>.279*</td>
<td>.471*</td>
<td>.346*</td>
<td>.087</td>
<td>.204</td>
<td>.142</td>
<td>.754*</td>
<td>.413*</td>
</tr>
<tr>
<td>Post10</td>
<td>.275*</td>
<td>.249*</td>
<td>.314*</td>
<td>.650*</td>
<td>.477*</td>
<td>.076</td>
<td>.367*</td>
<td>.248*</td>
<td>.532*</td>
<td>.757*</td>
</tr>
</tbody>
</table>

n=134, except on Pre10 and Post10 (Global Measure) where n=131
Note: * Correlations are significant at the p<.01 level. Dimension correlations are, in order, 1) breadth, 2) enthusiasm, 3) organization, 4) exams, 5) individual interaction, 6) workload/difficulty, 7) group discussion/interaction, 8) course demands/materials, 9) student performance/learning, and 10) a global measure of performance.

A closer examination of these correlations is done by correlating pre- and post-intervention item scores for the trained groups and the control group. Higher correlations were found on nine of the ten dimensions in the control group. A higher correlation on pre- and post-intervention scores in the control group is an accurate reflection of the lack of training. Because students in the control group received no manipulation following the initial performance rating, their evaluation scores should not have changed following the intervention. Trained students, on the other hand, should have a lower correlation with pre-intervention scores because an attempt to change their attitudes or ability to rate performance had been made. Trained student evaluation scores should correlate higher with the “true performance score” than with the pre-intervention score. Because the
Table 3
Correlations of scores taken before and after intervention for Trained Groups

<table>
<thead>
<tr>
<th></th>
<th>Pre1</th>
<th>Pre2</th>
<th>Pre3</th>
<th>Pre4</th>
<th>Pre5</th>
<th>Pre6</th>
<th>Pre7</th>
<th>Pre8</th>
<th>Pre9</th>
<th>Pre10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post1</td>
<td>.419*</td>
<td>.368*</td>
<td>.369*</td>
<td>.420*</td>
<td>.483*</td>
<td>-.040</td>
<td>.402*</td>
<td>.282*</td>
<td>.440*</td>
<td>.547*</td>
</tr>
<tr>
<td>Post2</td>
<td>.407*</td>
<td>.674*</td>
<td>.342*</td>
<td>.302*</td>
<td>.487*</td>
<td>.086</td>
<td>.445*</td>
<td>.311*</td>
<td>.351*</td>
<td>.463*</td>
</tr>
<tr>
<td>Post3</td>
<td>.255</td>
<td>.157</td>
<td>.757*</td>
<td>.417*</td>
<td>.277*</td>
<td>.254</td>
<td>.114</td>
<td>.095</td>
<td>.298*</td>
<td>.442*</td>
</tr>
<tr>
<td>Post4</td>
<td>.255</td>
<td>.126</td>
<td>.437*</td>
<td>.685*</td>
<td>.328*</td>
<td>.165</td>
<td>.136</td>
<td>.173</td>
<td>.589*</td>
<td>.475*</td>
</tr>
<tr>
<td>Post5</td>
<td>.295*</td>
<td>.449*</td>
<td>.236*</td>
<td>.397*</td>
<td>.692*</td>
<td>-.135</td>
<td>.331*</td>
<td>.296*</td>
<td>.445*</td>
<td>.473*</td>
</tr>
<tr>
<td>Post6</td>
<td>-.112</td>
<td>-.064</td>
<td>.295*</td>
<td>.102</td>
<td>-.170</td>
<td>.884*</td>
<td>-.096</td>
<td>-.089</td>
<td>.063</td>
<td>.039</td>
</tr>
<tr>
<td>Post7</td>
<td>.329*</td>
<td>.384*</td>
<td>.220*</td>
<td>.203</td>
<td>.389*</td>
<td>-.062</td>
<td>.600*</td>
<td>.235*</td>
<td>.320*</td>
<td>.424*</td>
</tr>
<tr>
<td>Post8</td>
<td>.238</td>
<td>.282*</td>
<td>.158</td>
<td>.362*</td>
<td>.518*</td>
<td>-.124</td>
<td>.257*</td>
<td>.594*</td>
<td>.395*</td>
<td>.457*</td>
</tr>
<tr>
<td>Post9</td>
<td>.329*</td>
<td>.212</td>
<td>.230*</td>
<td>.453*</td>
<td>.352*</td>
<td>.094</td>
<td>.134</td>
<td>.187</td>
<td>.709*</td>
<td>.416*</td>
</tr>
<tr>
<td>Post10</td>
<td>.315*</td>
<td>.245</td>
<td>.327*</td>
<td>.599*</td>
<td>.435*</td>
<td>.104</td>
<td>.351*</td>
<td>.224*</td>
<td>.524*</td>
<td>.710*</td>
</tr>
</tbody>
</table>

n=91 for all items except Post10 and Pre10 where n=88

Table 4
Correlations of scores taken before and after intervention for Control Group

<table>
<thead>
<tr>
<th></th>
<th>Pre1</th>
<th>Pre2</th>
<th>Pre3</th>
<th>Pre4</th>
<th>Pre5</th>
<th>Pre6</th>
<th>Pre7</th>
<th>Pre8</th>
<th>Pre9</th>
<th>Pre10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post1</td>
<td>.651*</td>
<td>.503*</td>
<td>.662*</td>
<td>.387</td>
<td>.356</td>
<td>.188</td>
<td>.434*</td>
<td>.191</td>
<td>.466*</td>
<td>.351</td>
</tr>
<tr>
<td>Post2</td>
<td>.576*</td>
<td>.880*</td>
<td>.566*</td>
<td>.323</td>
<td>.364</td>
<td>.105</td>
<td>.321</td>
<td>.116</td>
<td>.467*</td>
<td>.412*</td>
</tr>
<tr>
<td>Post4</td>
<td>.251</td>
<td>.235</td>
<td>.351</td>
<td>.713*</td>
<td>.412*</td>
<td>.248</td>
<td>.338</td>
<td>.198</td>
<td>.373</td>
<td>.577*</td>
</tr>
<tr>
<td>Post5</td>
<td>.360</td>
<td>.414*</td>
<td>.433*</td>
<td>.496*</td>
<td>.739*</td>
<td>-.135</td>
<td>.531*</td>
<td>.302</td>
<td>.585*</td>
<td>.466*</td>
</tr>
<tr>
<td>Post6</td>
<td>.206</td>
<td>.064</td>
<td>.426*</td>
<td>.274</td>
<td>-.164</td>
<td>.992*</td>
<td>-.005</td>
<td>.055</td>
<td>.100</td>
<td>.095</td>
</tr>
<tr>
<td>Post7</td>
<td>.178</td>
<td>.384</td>
<td>.292</td>
<td>.342</td>
<td>.730*</td>
<td>-.019</td>
<td>.850*</td>
<td>.504*</td>
<td>.252</td>
<td>.353</td>
</tr>
<tr>
<td>Post8</td>
<td>.201</td>
<td>.184</td>
<td>.296</td>
<td>.432*</td>
<td>.559*</td>
<td>.147</td>
<td>.514*</td>
<td>.899*</td>
<td>.324</td>
<td>.461*</td>
</tr>
<tr>
<td>Post9</td>
<td>.317</td>
<td>.419*</td>
<td>.386</td>
<td>.505*</td>
<td>.326</td>
<td>.087</td>
<td>.340*</td>
<td>.090</td>
<td>.876*</td>
<td>.393*</td>
</tr>
<tr>
<td>Post10</td>
<td>.149</td>
<td>.262</td>
<td>.282</td>
<td>.839*</td>
<td>.575*</td>
<td>.014</td>
<td>.403*</td>
<td>.328</td>
<td>.556*</td>
<td>.883*</td>
</tr>
</tbody>
</table>

n=43

Note: * Correlations are significant at the p<.01 level. Dimension correlations are, in order, 1) breadth, 2) enthusiasm, 3) organization, 4) exams, 5) individual interaction, 6) workload/difficulty, 7) group discussion/interaction, 8) course demands/materials, 9) student performance/learning, and 10) a global measure of performance.
control group was not trained, their ratings should be higher than the groups targeted by the training sessions.

**Correlating Student Grade with Evaluation Scores**

McKeachie (1979) states that "...courses in which students learn more the grades should be higher and the ratings (of teaching effectiveness) should be higher..." (p. 390-391). Student performance/learning may be attributed to effective teaching (Marsh, 1987). Thus, student grade was correlated with ratings of each of the ten dimensions of effective teaching. Dimension scores were correlated with grades for each student. Three measures of student grade were obtained. Each student's actual grade 11 weeks into the 16 week semester was obtained from the student's instructor. Additionally, each student was asked to report his or her current grade (to date) and expected final grade. All three measures of grade were significantly correlated (see Table 5).

<table>
<thead>
<tr>
<th></th>
<th>Grade</th>
<th>Reported Grade</th>
<th>Expected Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Reported Grade</td>
<td>.854**</td>
<td>----</td>
<td>---</td>
</tr>
<tr>
<td>Expected Grade</td>
<td>.781**</td>
<td>.878**</td>
<td>----</td>
</tr>
</tbody>
</table>

**Evaluation Score Descriptives**

Table 6 lists the mean and standard deviation for each dimension of teaching performance before and after the intervention as well as the instructor's self-evaluated
score. To examine the accuracy with which students evaluated their instructors, student evaluation scores were correlated with grades. Effective instruction should result in increased student learning (Day, 1995; Frey et al., 1975). Students who receive a high grade in the course may have learned more than those students with a lower grade. McKeachie (1979) stated that students who learn more should have a higher grade and likewise the student’s evaluation of teaching effectiveness will be higher. Thus, those students who received a high grade in the course should rate their instructor’s effectiveness higher. Correlations between student evaluations scores and student grade are represented in Table 7. Student grade correlated significantly with only three dimensions of performance. In two of the three cases the control group had a higher correlation with the dimension score than did either of the two trained groups. The lack of a significant, positive correlation with either of the two trained groups is not consistent with hypothesis one, which stated that the trained groups would be more closely related to student grade on each dimension of teaching. Although there are no significant differences between the two trained groups, consistent with hypothesis three, hypothesis one was not supported.

Comparing Student Evaluation Scores to Instructor Self-Evaluation Scores

Each item score was also compared with the instructor’s self-evaluation. A difference score was computed for each of the 10 evaluation items. The difference score was calculated by subtracting the student’s evaluation score from the instructor’s self-evaluation score. An absolute value difference score of zero would indicate the student
Table 6
Dimension Means and Standard Deviations

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Pre Measure</th>
<th></th>
<th>Post Measure</th>
<th></th>
<th>Self Measure</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Breadth</td>
<td>1.37</td>
<td>.53</td>
<td>1.40</td>
<td>.54</td>
<td>1.31</td>
<td>.46</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>1.35</td>
<td>.52</td>
<td>1.39</td>
<td>.56</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>Organization</td>
<td>1.55</td>
<td>.65</td>
<td>1.56</td>
<td>.63</td>
<td>1.31</td>
<td>.46</td>
</tr>
<tr>
<td>Exams</td>
<td>2.01</td>
<td>.88</td>
<td>1.99</td>
<td>.95</td>
<td>2.00</td>
<td>.00</td>
</tr>
<tr>
<td>Individual Interaction</td>
<td>1.74</td>
<td>.76</td>
<td>1.75</td>
<td>.74</td>
<td>1.74</td>
<td>.44</td>
</tr>
<tr>
<td>Workload / Difficulty</td>
<td>2.29</td>
<td>1.16</td>
<td>2.32</td>
<td>1.19</td>
<td>2.70</td>
<td>1.49</td>
</tr>
<tr>
<td>Group Discussion</td>
<td>1.72</td>
<td>.75</td>
<td>1.79</td>
<td>.73</td>
<td>2.00</td>
<td>.00</td>
</tr>
<tr>
<td>Course Demands / Materials</td>
<td>2.40</td>
<td>.96</td>
<td>2.34</td>
<td>.99</td>
<td>2.26</td>
<td>.44</td>
</tr>
<tr>
<td>Student Performance / Learning</td>
<td>1.97</td>
<td>.82</td>
<td>2.04</td>
<td>.83</td>
<td>2.31</td>
<td>.46</td>
</tr>
<tr>
<td>Global Measure</td>
<td>1.63</td>
<td>.71</td>
<td>1.65</td>
<td>.64</td>
<td>1.57</td>
<td>.50</td>
</tr>
</tbody>
</table>

n=134 on all items except item the Global Measure where n=131

agreed with the instructor. A larger absolute value of the difference score would indicate more disagreement between the student and the instructor.

A one-way analysis of variance was used to explore any differences between student evaluation score and instructor self-evaluation score across all three groups. The data in Table 8 indicate significant differences on only one dimension (Organization, 3.90, p<.05). A post-hoc analysis using Scheffe’s test of multiple comparisons identified that the difference was found between the lecture group and the control group, in support of hypothesis one. No significant differences were found between either of the two trained groups on any of the ten items, consistent with hypothesis four. The results of the ANOVA show that there are significant differences on only one dimension of performance, Organization (F=3.90, p<.05).
Table 7
Correlations with Grade

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Group</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breadth</td>
<td>Live</td>
<td>.208</td>
<td>.421**</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>-.083</td>
<td>.147</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>.227</td>
<td>.229</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>Live</td>
<td>.204</td>
<td>-.018</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>-.085</td>
<td>-.049</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>.164</td>
<td>.184</td>
</tr>
<tr>
<td>Organization</td>
<td>Live</td>
<td>.043</td>
<td>.039</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>-.120</td>
<td>.100</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>.028</td>
<td>.131</td>
</tr>
<tr>
<td>Exams</td>
<td>Live</td>
<td>.268</td>
<td>.076</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>.113</td>
<td>.076</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>.167</td>
<td>.317*</td>
</tr>
<tr>
<td>Individual Interaction</td>
<td>Live</td>
<td>.281</td>
<td>.086</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>.125</td>
<td>.078</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>.302*</td>
<td>.165</td>
</tr>
<tr>
<td>Workload / Difficulty</td>
<td>Live</td>
<td>-.030</td>
<td>.198</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>-.156</td>
<td>-.098</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>-.040</td>
<td>-.046</td>
</tr>
<tr>
<td>Group Discussion</td>
<td>Live</td>
<td>.062</td>
<td>-.104</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>.119</td>
<td>.023</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>.155</td>
<td>.142</td>
</tr>
<tr>
<td>Course Demands / Materials</td>
<td>Live</td>
<td>-.120</td>
<td>.200</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>.028</td>
<td>.143</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>.138</td>
<td>.027</td>
</tr>
<tr>
<td>Student Performance / Learning</td>
<td>Live</td>
<td>.492**</td>
<td>.226</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>.282*</td>
<td>.214</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>.476**</td>
<td>.332*</td>
</tr>
<tr>
<td>Global Measure</td>
<td>Live</td>
<td>.109</td>
<td>.040</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>-.120</td>
<td>-.138</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>.226</td>
<td>.111</td>
</tr>
</tbody>
</table>

n=134 on all items except Global Measure where n=131
Table 8  
One way ANOVA of difference scores across groups

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Test of Homogeneity</th>
<th>F value</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breadth</td>
<td>.099</td>
<td>6.17</td>
<td>.541</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>.081</td>
<td>1.53</td>
<td>.221</td>
</tr>
<tr>
<td>Organization</td>
<td>.005</td>
<td>3.90</td>
<td>.023</td>
</tr>
<tr>
<td>Exams</td>
<td>.375</td>
<td>2.17</td>
<td>.118</td>
</tr>
<tr>
<td>Individual Interaction</td>
<td>.230</td>
<td>.115</td>
<td>.892</td>
</tr>
<tr>
<td>Workload / Difficulty</td>
<td>432</td>
<td>.969</td>
<td>.382</td>
</tr>
<tr>
<td>Group Discussion</td>
<td>.892</td>
<td>.527</td>
<td>.592</td>
</tr>
<tr>
<td>Course Demands / Materials</td>
<td>.315</td>
<td>.912</td>
<td>.404</td>
</tr>
<tr>
<td>Student Performance / Learning</td>
<td>.870</td>
<td>1.31</td>
<td>.274</td>
</tr>
<tr>
<td>Global Measure</td>
<td>.143</td>
<td>1.11</td>
<td>.334</td>
</tr>
</tbody>
</table>

Examining the Global Measure of Performance

Day (1995) reported that a global measure of performance may account for a significant amount of variance. To explore this possibility, the global measure of performance was regressed on each dimension of performance. Table 9 displays the $R^2$ and significance of the variance accounted for by the global measure of performance. Only on the Workload / Difficulty dimension does the global measure of performance fail to account for a significant amount of variance ($R^2=.001$, $p<.05$). Additionally, grade does account for a significant amount of the variance on three of the ten items: Breadth ($R^2=.059$, $p<.01$), Exams ($R^2=.022$, $p<.05$), and Student Performance/Learning ($R^2=.081$, $p<.01$).
Table 9
Item variance accounted for by Global Measure and Grade.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Variable</th>
<th>R Square</th>
<th>F value</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breadth</td>
<td>Global</td>
<td>.225</td>
<td>37.38</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Grade</td>
<td>.059</td>
<td>10.59</td>
<td>.001</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>Global</td>
<td>.161</td>
<td>24.82</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Grade</td>
<td>.001</td>
<td>1.06</td>
<td>.745</td>
</tr>
<tr>
<td>Organization</td>
<td>Global</td>
<td>.167</td>
<td>25.83</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Grade</td>
<td>.000</td>
<td>0.24</td>
<td>.877</td>
</tr>
<tr>
<td>Exams</td>
<td>Global</td>
<td>.341</td>
<td>66.79</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Grade</td>
<td>.022</td>
<td>4.43</td>
<td>.037</td>
</tr>
<tr>
<td>Individual Interaction</td>
<td>Global</td>
<td>.320</td>
<td>60.64</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Grade</td>
<td>.014</td>
<td>2.74</td>
<td>.100</td>
</tr>
<tr>
<td>Workload / Difficulty</td>
<td>Global</td>
<td>.001</td>
<td>1.49</td>
<td>.700</td>
</tr>
<tr>
<td></td>
<td>Grade</td>
<td>.016</td>
<td>2.06</td>
<td>.154</td>
</tr>
<tr>
<td>Group Discussion</td>
<td>Global</td>
<td>.217</td>
<td>35.68</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Grade</td>
<td>.001</td>
<td>1.18</td>
<td>.672</td>
</tr>
<tr>
<td>Course Demands / Materials</td>
<td>Global</td>
<td>.197</td>
<td>31.69</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Grade</td>
<td>.010</td>
<td>1.65</td>
<td>.202</td>
</tr>
<tr>
<td>Student Performance / Learning</td>
<td>Global</td>
<td>.341</td>
<td>66.61</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Grade</td>
<td>.081</td>
<td>17.85</td>
<td>.000</td>
</tr>
</tbody>
</table>

Discussion

Higher intra-dimensional correlations on pre-/post-intervention evaluation scores is a clear indication students are able to discriminate between performance dimensions. However, this trend was not consistent with the present hypotheses. Because the intra-dimensional correlations for the trained groups do not significantly differ from the control group, no judgements can be made regarding the effectiveness of the training. Significant correlations between the trained groups and the instructor’s self-evaluation scores might have illustrated an effective attempt to train student raters.
To evaluate the accuracy with which students evaluated their professors, student evaluation scores were correlated with student grade and instructor self-evaluation. McKeachie (1979) stated that students who learn more in the course will receive a higher grade and will rate their instructor as being more effective. Three measures of student grade were collected: actual grade, self-reported grade, and expected final grade. The student’s actual grade was the grade obtained from the instructor 11 weeks into the 16 week semester. All three measures of student grade were significantly correlated. A high correlation between all three measures of student grade suggests that students accurately report their grade in the course. To avoid any unnecessary potential problems with inaccurate grade reporting by the students, the student’s actual grade (obtained from the instructor) was used to assess evaluation accuracy in the present study.

Trained students should assess their instructor’s effectiveness more accurately than students not trained. Correlations between student grade and student evaluation score should have been higher in the trained groups than in the control group. Trained students should have been better able to distinguish between effective and ineffective classroom performance. However, because no significant correlations could be found between student ratings and grade hypotheses one and two were not supported by the present data.

Student evaluation scores were also compared to the instructor’s self-evaluation. Support for the validity of instructor self-evaluations is well-documented (Marsh, 1987, Schneider et al., 1994). Given an accurate evaluation of classroom performance from the instructor, accurate student evaluations should correlate highly with instructor self-evaluations. No significant differences were found between student and instructor
evaluation scores. Low agreement between student and instructor could mean that one or both of the evaluations are inaccurate.

Day (1995) found that a global measure of performance accounted for a significant amount of variance on each dimension of teaching. Despite the multidimensionality of performance, when assessing the overall performance of the instructor, a global measure of performance accounts for the most variance. When judging the overall performance of an instructor, a global measure of performance is just as useful as multiple dimension scores. Dimensional performance scores would be most helpful when evaluating performance for purposes of feedback. Feedback should be specific to allow for performance improvement. Global measures of performance do not allow for such feedback specificity. However, judgements of performance (e.g., promotions, tenure, etc.) are not made by judging performance on specific dimensions. Professors are not granted tenure because they are organized or because they give good examinations. Rather, decisions are made by judging overall performance. Because the global measure of performance is able to account for a significant amount of variance on performance scores for each dimension, the global measure of performance may well be the most useful rating of effective classroom performance.

Grade is able to account for a significant amount of variance on three dimensions of performance: breadth, student performance/learning, and exams. A logical argument can be made for the relationship between grade and each of these three items. Students who have effectively learned the material presented in class will receive a higher grade in the course. Additionally, students subjected to a large breadth of material in class are
likely to have learned more about the subject in class. Finally, in most classrooms, student grade is comprised mainly of exam scores during the semester. It is not surprising that grade accounts for a significant amount of variance on each of the three dimensions of teaching performance.

There are several limitations of the present study that should be noted. Some methodological concerns to consider include the categorization of student grade. Students were assigned to one of five categories for grades (100-90, 89-80, 79-70, 69-60, and 59 and below). Student evaluation scores were correlated with the student grade operationalized as a one through a five. The student grade could have been represented as a continuous variable one through 100. A continuous grade variable would allow for greater variation in student grade and may have accounted for more variance in student evaluation scores.

Another potential problem is the use of only one item per dimension to measure teaching effectiveness. Additional items measuring performance on the same dimension may have improved the reliability of the measures because we would have more than one item measuring performance in the same dimension. Additional items measuring performance on each dimension should increase the reliability and validity of the student’s measure of instructional performance.

The use of student grade as a criterion, in the present study, to judge student evaluation accuracy may not have been the most accurate. Student grade was chosen as a criterion measure from McKeachie (1997). McKeachie stated that grades may have an impact on student evaluation outcomes. The impact of student grade on evaluation
outcome is not questioned. However, controlling for student grade in group assignment, as was done in the present study, was a sufficient exercise to address grade leniency concerns.

An additional problem may have been the sample size of the study. Approximately 40-50 students made up each group. An ideal sample size would have been closer to 80-100 for each group. The present sample was taken from available Introductory Psychology courses in the Spring semester. The lack of large section classes limited the available sample for of this study. Courses were chosen on the basis of teaching multiple sections. Only three professors agreed to participate, leaving a potential sample size of 200. Collecting data for this study in the Fall semester when there are more large section classes available would greatly facilitate data collection.

Finally, the brevity of the training program presents a potential problem. The present training program was designed to be administered within the time parameters of a single class period (one hour). It was necessary to administer the student training as well as collect all student evaluation measures within a single class period. The present training program lasted approximately 10 minutes. Past training programs have included dimensional and error training between one hour (Bernardin, 1978) and up to 14 hours (Ivancevich, 1979). The majority of training programs discussed in this review of literature were more than one hour in length (Smith, 1986). The treatment in the present study may not have been effective enough to improve the students' ability to rate their instructor's classroom performance.
In conclusion, the present study was an attempt to improve the accuracy of student evaluation scores through dimension and frame-of-reference training. Rater training was presented via two media: lecture-based training and video-taped training. Although students were able to discriminate between dimension items, student evaluation scores for the trained groups did not differ significantly from the control group.
References


Crittenden


Appendix A

STUDENT RATER TRAINING

I am Dr. Betsy Shoenfelt. I am an industrial/organizational psychologist on the faculty in Western’s Psychology Department. I am here today to provide you with some information on how to make accurate and fair ratings of faculty teaching.

The reason I am here is two-fold:

1. Teaching is an important component of a faculty member’s job. Your assessment of your instructor’s teaching is frequently the only information a department head has on classroom performance. This means your evaluation is an important component of the faculty member’s annual evaluation.

2. It is sometimes difficult to accurately evaluate teaching performance, or any performance, for that matter. In fact, training raters is one of the most important steps in increasing the accuracy and fairness of ratings.

Many organizations recognize this fact and provide training for managers and supervisors who complete performance evaluations. I have done training similar to the training you will receive here today for groups such as: the United States Department of Defense, Kentucky State Department of Education, General Electric, Eaton Axle & Brake, the Kendall Company, Batesville Casket, Camping World, as well as for the deans and department heads at Western.

This training will take approximately 15 minutes. It should help ensure that the ratings you give your course instructor will be accurate and fair. In addition, the information you will hear in this brief training session will be useful to you if you are employed and either have your own performance evaluated or have a job as a manager or supervisor where you will need to evaluate the performance of individuals working for you.

One of the most important things to recognize when completing teaching evaluations is that there are a number of different types or dimensions of performance that are necessary for effective teaching. The same instructor may be very good in some aspects of teaching, only average in other aspects of teaching, and poor in still other dimensions of teaching. All of us have strengths and weaknesses and these should be reflected in our evaluations.

For example:

A teacher who shows excitement when answering questions in class would be considered to be very high on Enthusiasm. The same individual might be average
on the Difficulty of the Work Load Assigned to the class, and poor on Exams, for example, failing to grade and return exams that students have taken.

Another teacher might not be as Enthusiastic but be very good at constructing exams and providing students with feedback from the exams.

It is sometimes tempting as a rater to give all favorable ratings because the instructor is nice or because we like the instructor. It is as though there is a “halo” above this instructor’s head, so we are inclined to give high, favorable ratings on all of the items on the rating form. We can also have negative halo if for some reason we have an unfavorable impression of the instructor and we give negative ratings across all of the items on the rating form.

However, the only fair way to rate any instructor is based on performance, in (and sometimes outside of) the classroom, on the dimensions of performance identified on the rating scale.

Another common error in evaluating performance is to be overly lenient or overly harsh. With leniency, usually the rater is reluctant to give anybody a “bad” score, so he or she gives easy or high ratings that really are not deserved by the instructor’s performance. With severity, usually the rater feels that he or she has such higher personal standards that nobody can measure up to these standards. In this case, the rater gives ratings that are lower than the instructors deserve. In either case, the ratings are not accurate or fair because they do not reflect the instructor’s performance.

Again, the important principle is to think in terms of the behavior you have seen your instructor demonstrate and compare that behavior to standards of performance for each dimension of teaching. Then give the instructor the rating for that dimension that reflects the performance you have seen.

**DIMENSIONS OF FACULTY TEACHING**

To illustrate this further, let’s look at nine important dimensions of university teaching and example of good and poor performance in each dimension:

**1. BREADTH**

- The instructor gives background of ideas/concepts, gives different points of view, discusses current developments, and displays a clear understanding of course topics.

- **GOOD:** The instructor touched on many aspects of the topic and explained that there was more to the topic that s/he presented, providing an overview without overwhelming the students.
POOR: The instructor presented isolated theories without indicating their relevance to the other theories or topics.

2. ENTHUSIASM - The instructor displays enthusiasm about teaching, is dynamic and energetic, enhances presentations with humor, demonstrates a teaching style that holds the interest of the class, and shows interest in teaching the class.

GOOD: The instructor demonstrated enthusiasm for the material which helped to maintain the interest of the students.

POOR: The instructor seemed bored with his/her own lecture.

3. ORGANIZATION - The instructor uses clear explanations, prepares course materials, states and follows course objectives, has well organized lectures and is well prepared for class.

GOOD: The instructor used visual aids and diagrams that clarified and explained difficult subject matter.

POOR: The instructor failed to provide a detailed syllabus for the course.

4. EXAMS - Performance measures (exams, assignments, etc.) are well constructed, the exams provide valuable feedback; tests emphasize course content.

GOOD: The instructor graded exams in a timely fashion and returned them quickly so that students were always aware of their standing in class.

POOR: The instructor gave exams that covered material that was not included in the assigned readings, handouts, or lectures.

5. INDIVIDUAL INTERACTION - The instructor is actively helpful, friendly toward students, welcomes seeking advise, is interested in individual students, and is accessible to individual students.

GOOD: The instructor stayed after class to tutor students who were having difficulty.

POOR: The instructor did not return e-mail messages from students.

6. WORKLOAD/DIFFICULTY - The instructor assigns an appropriate amount of work of an appropriate difficulty level.
GOOD: Assignments were challenging but no insurmountable.

POOR: The instructor assigned many projects that were “busy work” and did not contribute to student understanding.

7. GROUP DISCUSSION/INTERACTION - Students are encouraged to ask questions and express their ideas. The instructor responds with meaningful answers.

GOOD: During the lecture the instructor encouraged students to ask questions and make comments.

POOR: The instructor ridiculed students for wrong answers.

8. COURSE DEMANDS/MATERIALS - Readings and texts are valuable and contribute to course understanding; assignments (readings, homework, etc.) contribute to the appreciation and understanding of the subject.

GOOD: Homework problems are assigned to illustrate points covered in class. Problems are practical and challenging.

POOR: Homework assignments are given simply to “take attendance” and are not related to the objectives of the course.

9. STUDENT PERFORMANCE/LEARNING - Students understand/learn the subject material in the course and increase their interest in the subject matter.

GOOD: The student is able to take information learned in class and apply it in his/her own life.

POOR: The student failed to read assigned material and was unable to understand the material presented in class.

Remember, these are examples of good and poor performance within each dimension. Your instructor may not have demonstrated these exact behaviors in each dimension, but he or she likely demonstrated behavior that is comparable.

Again, the important principle is to think in terms of the behavior you have seen your instructor demonstrate and compare it to the standards of performance for each dimension of teaching. Then give the instructor the rating for that dimension that reflects the
performance you have seen. Also bear in mind that most instructors have strengths and weaknesses and that should be reflected in your performance ratings.

The ratings you give your instructor are an important part of his or her annual evaluation. Please do your best to give fair and accurate ratings.

Thank you.
Appendix B

General Introduction

Proctor says: “Hello! I would like to begin this morning by thanking all of you for participating. My name is Rob Beeler. I am a graduate student in the Psychology Department conducting a study as part of my Master’s degree requirements. We are studying student ratings of college teaching performance. The evaluations you will perform will be very similar to the evaluations you do at the end of each semester.”

Proctor says: “We will be dividing the class into three different groups and taking each group into a separate room. Each group will use a slightly different procedure to complete the ratings. Previous research indicates that all three of the procedures we will be using are effective for teaching evaluation. You will be told the details of the procedure you will be using after you are in your research group.”

Proctor says: “Participation in the research is voluntary, however, if you choose not to participate you will still need to remain in the assigned room until your regular class period is over.”

Assigning Rooms

A class roster should be provided that lists all students as A, B, C, or below. One “A” student will be assigned to Group One followed by one “A” student to Group Two and so on until all students have been assigned. If students are absent from class, simply pass them over.

Proctor says:

“The following students are assigned to Group One. You will go with ____ to room ____.”

“The following students are assigned to Group Two. You will go with ____ to room ____.”

“The following students are assigned to Group Three. You will go with ____ to room ____.”
Group One: Lecture Training

Pass out consent forms.

Proctor says: “Participation in this study is voluntary. It is important that you know you are at no risk by agreeing to participate in this study. Please read over the consent form you have in front of you and sign it.”

Proctor says: “The consent form lists the nature and purpose of the experiment. It also explains that there are no risks to you and that your responses are for research purposes only and will be kept confidential.”

After all participants have had a chance to sign, ask them to pass the consent forms forward.

Pass out Instrument A rating forms and scantrons.

Proctor says: “We would first like you to complete these ratings just as you would the official university ratings that are given toward the end of any semester in every course. DO NOT put your name on the answer sheet. Your responses will be anonymous. However, the researchers are interested in whether there are differences between groups of respondents. For example, do males and females differ in their teacher ratings. Therefore, we need some biographical information. Your responses should be recorded only on the scantron sheet. Please do not write on the rating instrument.”

Proctor says: “First, please write your instructor’s name on the scantron where it asks for NAME. Next, write your course number (e.g., PSY 100) on the scantron where it asks for SUBJECT. In the TEST NO. block, write your age. For question one, fill in A if you are a male and B if you are a female. For question two, fill in A for White, B for Black, C for Hispanic, D for Asian, and E for other. For question three, fill in the final grade you expect to receive in this course (A for A, B for B, C for C, D for D, and E for below 60). Question four asks for your current grade in this course based on exams and assignments to date. Please darken the appropriate item on the scantron. For the remainder of the items use the following scale:

A = Strongly Agree
B = Agree
C = Undecided
D = Disagree
E = Strongly Disagree
Proctor says: “Do not write on the rating instrument. Mark all of your answers on the scantron.”

Give class time to complete evaluations.

Collect Rating Instrument A.

Dr. Shoenfelt administers lecture training session.

Pass out Rating Instrument B.

Proctor says: “Because evaluations are so important, we would like you to be especially careful in completing your evaluations to ensure that they are accurate and fair ratings.”

Give class time to complete evaluations.

Collect Instrument B.

Proctor says: “You will be asked to complete another evaluation of your professor’s performance this semester. Are there any questions? Thank you for your time.”
Group Two: Video Tape Training

Pass out consent forms.

**Proctor says:** “Participation in this study is voluntary. It is important that you know you are at no risk by agreeing to participate in this study. Please read over the consent form you have in front of you and sign it.”

**Proctor says:** “The consent form lists the nature and purpose of the experiment. It also explains that there are no risks to you and that your responses are for research purposes only and will be kept confidential.”

After all participants have had a chance to sign, ask them to pass the consent forms forward.

Pass out Instrument A rating forms and scantrons.

**Proctor says:** “We would first like you to complete these ratings just as you would the official university ratings that are given toward the end of any semester in every course. DO NOT put your name on the answer sheet. Your responses will be anonymous. However, the researchers are interested in whether there are differences between groups of respondents. For example, do males and females differ in their teacher ratings. Therefore, we need some biographical information. Your responses should be recorded only on the scantron sheet. Please do not write on the rating instrument.”

**Proctor says:** “First, please write your instructor’s name on the scantron where it asks for NAME. Next, write your course number (e.g., PSY 100) on the scantron where it asks for SUBJECT. In the TEST NO. block, write your age. For question one, fill in A if you are a male and B if you are a female. For question two, fill in A for White, B for Black, C for Hispanic, D for Asian, and E for other. For question three, fill in the final grade you expect to receive in this course (A for A, B for B, C for C, D for D, and E for below 60). Question four asks for your current grade in this course based on exams and assignments to date. Please darken the appropriate item on the scantron. For the remainder of the items use the following scale:

A = Strongly Agree
B = Agree
C = Undecided
D = Disagree
E = Strongly Disagree
**Proctor says:** “Do not write on the rating instrument. Mark all of your answers on the scantron.”

Give class time to complete evaluations.

Collect Rating Instrument A.

**Proctor says:** “The video you are about to watch is a training session by Dr. Betsy Shoenfelt. The training is designed to help you provide fair and accurate ratings of your instructor. After the video you will be asked to rate your professor on nine dimensions of performance discussed by Dr. Shoenfelt in the video. Before we begin are there any questions?”

**Proctor says:** “Please pay careful attention to the video.”

After video, pass out Rating Instrument B.

**Proctor says:** “Because evaluations are so important, we would like you to be especially careful in completing your evaluations to ensure that they are accurate and fair ratings.”

Give class time to complete evaluations.

Collect Instrument B.

**Proctor says:** “You will be asked to complete another evaluation of your professor’s performance this semester. Are there any questions? Thank you for your time.”
Group Three: Control Group

Pass out consent forms.

Proctor says: “Participation in this study is voluntary. It is important that you know you are at no risk by agreeing to participate in this study. Please read over the consent form you have in front of you and sign it.”

Proctor says: “The consent form lists the nature and purpose of the experiment. It also explains that there are no risks to you and that your responses are for research purposes only and will be kept confidential.”

After all participants have had a chance to sign, ask them to pass the consent forms forward.

Pass out Instrument A rating forms and scantrons.

Proctor says: “We would first like you to complete these ratings just as you would the official university ratings that are given toward the end of any semester in every course. DO NOT put your name on the answer sheet. Your responses will be anonymous. However, the researchers are interested in whether there are differences between groups of respondents. For example, do males and females differ in their teacher ratings. Therefore, we need some biographical information. Your responses should be recorded only on the scantron sheet. Please do not write on the rating instrument.”

Proctor says: “First, please write your instructor’s name on the scantron where it asks for NAME. Next, write your course number (e.g., PSY 100) on the scantron where it asks for SUBJECT. In the TEST NO. block, write your age. For question one, fill in A if you are a male and B if you are a female. For question two, fill in A for White, B for Black, C for Hispanic, D for Asian, and E for other. For question three, fill in the final grade you expect to receive in this course (A for A, B for B, C for C, D for D, and E for below 60). Question four asks for your current grade in this course based on exams and assignments to date. Please darken the appropriate item on the scantron. For the remainder of the items use the following scale:

A = Strongly Agree
B = Agree
C = Undecided
D = Disagree
E = Strongly Disagree

Proctor says: “Do not write on the rating instrument. Mark all of your answers on the scantron.”
Give class time to complete evaluations.

Collect Rating Instrument A.

Pass out Rating Instrument B.

**Proctor says:** “At this time we are going to ask you to complete a second evaluation of your instructor. This set of items is similar to those items you just completed, but because evaluations are so important, we would like you to be especially careful in completing your evaluations to ensure that they are accurate and fair ratings. Your ratings may be similar to the ones you just completed or you may feel you would like to change some of your ratings. Please be careful to give fair and accurate ratings. Are there any questions?”

Give class time to complete evaluations.

Collect Instrument B.

**Proctor says:** “You will be asked to complete another evaluation of your professor’s performance this semester. Are there any questions? Thank you for your time.”
Appendix C

Rating Instrument A

DO NOT put your name on the answer sheet. Your responses will be anonymous. However, the researchers are interested in whether there are differences between groups of respondents. For example, do males and females differ in their teacher ratings. Therefore, we need some biographical information. Your responses should be recorded only on the scantron sheet. DO NOT write on this form.

Scantron Block You write...
NAME Instructor's name
SUBJECT Course number
TEST NO. Your age
HOUR Group number

Answer the following questions on your scantron answer sheet.

1. What is your sex? A = Male  
   B = Female

2. What is your race? A = White  
   B = Black  
   C = Hispanic  
   D = Asian  
   E = Other

3. What is your expected final grade in this course? (A for A, B for B, etc.)

4. Based on exams and assignments thus far, what is your current grade in this course?

For each item listed below, carefully read the item then rate your instructor according to the following scale.

   A = Strongly Agree  
   B = Agree  
   C = Undecided  
   D = Disagree  
   E = Strongly Disagree

5. Performance measures (exams, assignments, etc.) are well-constructed.
6. My instructor is actively helpful.
7. I have learned and understood the subject materials in this course.
8. Readings, homework, etc. contributed to the appreciation and understanding of the subject.
9. Relative to other courses, the course workload was light.
10. My instructor is well-prepared for class.
11. My instructor displays interest in teaching this class.
12. Students were encouraged to ask questions and were given meaningful answers.
13. My instructor displays a clear understanding of course topics.
14. Overall, my instructor is effective.

When you have completed your ratings, please remain seated until all others have finished. Thank you!
Rating Instrument B

For each item listed below, carefully read the item then rate your instructor according to the following scale.

A = Strongly Agree
B = Agree
C = Undecided
D = Disagree
E = Strongly Disagree

15. My instructor displays a clear understanding of course topics.
16. My instructor displays interest in teaching this class.
17. My instructor is well-prepared for class.
18. Performance measures (exams, assignments, etc.) are well-constructed.
19. My instructor is actively helpful.
20. Relative to other courses, the course workload was light.
21. Students were encouraged to ask questions and were given meaningful answers.
22. Readings, homework, etc. contributed to the appreciation and understanding of the subject.
23. I have learned and understood the subject materials in this course.
24. Overall, my instructor is effective.

When you have completed your ratings, please remain seated until all others have finished.
Thank you!