



A Framework for Training Students as Evaluators of Instructor Performance

Linda S. Hartenian
University of Wisconsin Oshkosh

ABSTRACT

Student evaluations of instructor performance are important tools for instructor development and assessment. A framework for training students as evaluators of instructors is presented, incorporating four themes from the performance evaluation research—rating errors, rater accuracy, cognitive processes, and tangential factors. Goals and methods for the training program, as well as administrative issues, are presented. Finally, evaluation of benefits and costs of the training program is discussed.

A Framework for Training Students as Evaluators of Instructor Performance

Instructors typically undergo periodic evaluations of their teaching performance in conjunction with university or college policies. While purposes of these evaluations can differ (e.g., tenure, promotion, merit, certification, development of teaching skills), students often are given an opportunity to provide input about instructor performance in the classroom. Hence, they play an important role in the instructor development and evaluation process. Student involvement makes sense for a couple of reasons (Tuckman, 1995): a) students have an opportunity to regularly observe instructors in class; and, b) students are customers of the university and should provide feedback on how well instructors perform (also see Schneider, Hanges, Goldstein, & Braverman, 1994). Though some might feel that students are products (rather than customers) of educational systems, this article errs on the side of involvement of all constituencies!

A critical assumption is that the students completing evaluations have some degree of precision and skill in performing this task. Researchers suggest that this is not the case--student evaluations are fraught with accuracy problems (cf. Nyirenda, 1994). Despite shortcomings of student evaluations, universities continue to treat student evaluations as important assessment tools (Greenwald, 1997; Greenwald & Gilmore, 1997a). Yet, a literature search revealed no articles directed toward training student evaluators. At a minimum, having discussions with students about their role in instructor evaluation would be an important step in improving the process (Smith, 1986). The argument is developed below that universities should go further by developing a unified approach toward training students.

Four themes emerging from a review of the literature are incorporated into a unified framework for training student evaluators—rater/student errors (biases) in evaluation, rater/student accuracy, rater cognitive processes, and tangential factors that affect student judgments (see Bernardin & Walter, 1977; Cronbach, 1977; and Marsh, 1984). Administrative issues in designing and implementing a training program are included. Finally, the evaluation of costs and benefits of such a program is discussed.



A Training Program for Student Evaluators

While one best way to train students does not exist, taking a unified approach toward student training is useful if we are to address the multiple issues raised in the literature review. The design of the training program begins with identification of the Goals for training. Four goals are presented which are designed to improve the accuracy of student evaluations. Four training modules are created to reflect each Goal. Several issues are considered simultaneously in the design of each training module (Gagne & Briggs, 1979): the need for training; the trainee's cognitive, emotional, and behavioral processes; and, the basic tenets of learning (refer to the Appendix for a description and general examples of events of instruction as they apply to any of the Goals presented below).

Goal 1: Understanding Dimensions of Instructor Performance

Goal 2: Providing Fair Evaluations

Goal 3: Understanding the Broader Context of Instructor Evaluations

Goal 4: Preparing for the Evaluation Process

These goals are presented in the order in which training sessions might be conducted. Goals 1, 2, and 3 could be addressed during a new student orientation. Because of the importance of individual feedback to trainees, a workshop format for training is suggested in Goals 1, 2, and 3. Given resource constraints, designers of the student training may find that lecture and discussion with general feedback will serve to improve accuracy (Athey & McIntyre, 1987). Goal 4 is directed toward returning students and should be conducted in smaller discussion groups. This allows students to process their actual evaluation experiences after training in Goals 1, 2, and 3 (Bargh & Schul, 1980).

Before beginning Goal 1 training, students could complete a short questionnaire on attitudes toward contaminants in the rating process. For example, students might be asked what they think of an instructor who assigns more than the average amount of work during a semester, what kinds of grades they expect in courses given their current grade point average, and if they are likely to give lower ratings to an instructor they dislike. Questionnaire data are held for Goal 2 discussion.

Goal 1: Understanding Dimensions of Instructor Performance

The purpose of Goal 1 is to provide an opportunity for the student to generate a cognitive schema for performance evaluations. Knowledge of dimension names and definitions is a prerequisite to developing the student's skills in the cognitive processes of observing, storing, and recalling information and rating an instructor's performance. Links have been demonstrated between these processes and accuracy (DeNisi & Williams, 1988). For example, Bernardin and Walter (1977) found that students who were trained in how to use an evaluation form exhibited fewer rater errors. Day and Sulsky (1995) found that proper categorization resulted in more accurate ratings. Also called frame-of-reference training, defining dimensions and teaching proper categorization of instructor behaviors provides raters with empirically developed standards for performance (Hedge & Kavanaugh, 1988). Training also increases the likelihood that information will be accessible (i.e., recalled) when the time comes to complete the instructor



evaluation (Woehr, 1992). In summary, student raters should become familiar with the “job” they are being asked to evaluate (Heneman, Wexley, & Moore, 1987). Table 1 summarizes definitions of rater errors and provides recommendations for correcting rating errors. Examples of dimensions, terms, and measurement scales are suggested below. Specific training guidelines are then presented (refer to Table 2).

Table 1 Rater Errors: Definitions and Recommended Strategies for Correction	
<u>Definitions</u>	<u>Methods for Correcting</u>
<u>Leniency</u> : All instructors rated high	For leniency, severity and central tendency: a) Recognize performance variability b) Be fair in evaluations c) Define dimensions/use behavioral anchors
<u>Severity</u> : All instructors rated low	
<u>Central Tendency</u> : All instructors rated average	
<u>Halo</u> : Instructor rated high (average/low) on all dimensions because instructor is high (average/low) on one dimension	Reinforce that dimensions of performance are mutually exclusive and independent of one another
<u>First Impression</u> : Early attitudes toward instructor determine ratings at end of semester	Take notes during semester (i.e., diary)
<u>Recency Effect</u> : Recent behavior is weighted more heavily than earlier behaviors	Take notes during semester (i.e., diary)
<u>Contrast Effect</u> : Knowledge of previous performance levels (or others' performance) influences ratings in presents situation	Define dimensions/use behavioral anchors
<u>Similar-to-me Effect</u> : Instructor with qualities like student is rated more highly.	Provide instructor's [job] description, including performance expectations

Table 2 Recommendations for Goal 1: Understanding Dimensions of Instructor Performance
<u>Learning Objectives</u> : To accurately define dimensions, terms, and standards (levels) of performance



To correctly place instructor behaviors into dimensions

To identify levels of instructor performance

Workshop Format:

Demonstration, Discussion, Practice, Feedback, Discussion

Video Segment 1:

Demonstration, Discussion, Practice, Feedback, Discussion

Video Segment 2:

Instructor Behaviors Varied by Level of Performance Across Dimensions:

Learning to Distinguish Performance Levels

Suggestions:

Provide clear standards (expectations) for instructor performance

Use behaviorally-based evaluation measures

Use video segments that reflect actual evaluation settings

Dimensions

Dimensions of performance must be identified and defined. Dimensions are broad conceptual categories for describing what instructors do in the classroom, such as delivery of instruction, student/instructor interaction, evaluation techniques, and classroom management (see Tuckman, 1995, and Nyirenda, 1994). Feldman (1989) concluded that student ratings could represent as many as 28 different dimensions; research continues to explore the dimensionality of instructor ratings (Marks, 2000). Whether a university or college adopts dimensions that others have created or chooses to create its own, dimensions should be comprehensive and mutually exclusive (Binning & Barrett, 1989). (The term “university” will be subsequently used.)

Dimensions should be defined. The student/instructor interaction dimension, for example, might be defined as “the extent to which the instructor maintains effective communication with the students, is aware of the students’ developmental and emotional characteristics, shows compassion and empathy, and sincerely wants students to learn.” Once the dimensions are defined, more specific behavioral examples of how an instructor demonstrates performance are provided. To continue with the above example, the following behaviors might represent the student/instructor dimension: a) this instructor provides extra help to students who request it, b) this instructor praises or encourages students when they give a correct response, c) when students make comments, their contributions are accepted without disagreement or further discussion, d) this instructor corrects a student’s incorrect response in a condescending manner, and e) this instructor does not use the student’s name when addressing him or her. Note that behaviors are positively and negatively phrased; however, as indicated below (see Measurement Scales),



students are not asked to evaluate the meaning of the behaviors. Consider the last behavioral example. In a lecture room with 300 students, not referring to students by name may be more understandable (i.e., permissible) than in a class of 20 students.

Terms

Specific terms also should be defined. While some terms are familiar, others are likely to be strange to students and others: constructs, dimensions of performance, criteria, and standards [levels] of performance. Terms may be part of the process itself (e.g., the formal name of the evaluation form). Be sure to clarify if terms have specific applications. For example, the term “instructor” may refer to a professor or to a teaching assistant who leads a discussion section of a larger class. Once terms and dimensions have been defined, measurement scales are created.

Measurement Scales

A measurement scale represents levels (i.e., standards) for evaluating an instructor’s performance. Measurement scales can have as many levels as desired, though five or seven are recommended. The following examples demonstrate 5-level scales: a) below average, slightly below average, average, slightly above average, above average; b) below expectations, slightly below expectations, meets expectations, slightly above expectations, exceeds expectations; c) poor, acceptable, good, very good, excellent. Regardless of which terminology is used or how many levels are desired, each level should be defined. The process of defining levels is referred to as “providing anchors” for the standards of performance.

As mentioned earlier, behavioral examples of what students might expect an instructor to exhibit should be created. A substantive amount of research exists to support the use of behavioral definitions for the anchors and behavioral examples of instructor performance (e.g., Hartel, 1993). Behavioral-based ratings are more accurate (cf. Weirsmas, VandenBerg, & Latham, 1995); use of personality traits can result in rater errors (cf. Borman & Dunning, 1975). When ratings are based on clear standards and observable information, they are less susceptible to interpersonal affect (Park & Sims, 1989).

Recognize that the above recommendations to use behavioral-based methods assume that students are being asked to record their observations of instructor behaviors. A more controversial issue (and one beyond the scope of this article) is whether or not to have students interpret and judge the meaning of those behaviors. One way to minimize intentional distortion of ratings is for students to record frequency of instructor behaviors; use a student evaluation team to provide periodic feedback to the instructor, with final interpretation and judgment about instructor performance coming from the instructor (self-assessment) and peers (peer assessment).

Training Program

Learning objectives for Goal 1 should be put in writing and shared with students in a handout (see Table 2). Training begins with students viewing videos of instructors in actual classroom settings. Videos are recommended rather than written descriptions (i.e., “paper people”) because they result in better accuracy (Ryan, Daum, Bauman, Grisez, Mattimore,



Nalodka, & McCormick, 1995). Observed behavior results in better retention and retrieval of training material from memory (Kinicki, Hom, Trost, & Wade, 1995) and provides multiple cues (e.g., visual, auditory), which may result in deeper processing of information (Murphy & Balzer, 1986).

A carefully crafted video will include several demonstration, discussion, and practice segments. Begin with a demonstration of instructor behaviors that represent each dimension (Segment 1). Following the demonstration segment, the trainer allows for discussion of the observed behaviors and the dimensions used on the evaluation form. Students then practice evaluating a video segment that includes a variety of instructor behaviors. The trainer presents correct conclusions (i.e., identification of behavior and placement of behavior in a dimension) and allows for discussion of the practice segment. The next demonstration segment presents several instructor behaviors, but varies performance levels (Segment 2). The trainer discusses with students the examples of various instructor behaviors and why the levels of performance differ in each example. Students are presented with the practice segment in which they are asked to rate instructor behaviors using the measurement scale. The trainer provides feedback on and allows for discussion of the conclusions regarding the observed levels of performance.

In concluding the training under Goal 1, students are reminded that they are expected to recall this training when they proceed to Goals 2, 3, and 4 and in the actual evaluation setting. Once students have developed a common terminology and understanding, they are ready to begin the more challenging training--learning to evaluate consistently.

Goal 2: Providing Fair Evaluations

The term “fairness” frequently is divided into two types--procedural (process) and distributive (outcome). For the purposes of Goal 2, fairness of student evaluations is a procedural issue for two reasons. First, student evaluations are inputs into evaluation and development decisions about an instructor. They are interpreted by an instructor’s peers and/or administration to arrive at an outcome (e.g., merit pay, identifying a need for instructor development). Second, Goal 2 is concerned specifically with the process that students follow when completing instructor evaluations. If instructors are to believe in the fairness of the performance evaluation system that includes student evaluations, then instructors must perceive that students will provide fair evaluations. Universities should be concerned about instructor perceptions of fairness--Harris (1988) suggested that perceptions of unfairness lead to decreased performance.

One way to influence these perceptions would be to ensure that student evaluations were reliable and accurate. Training improves the reliability of ratings by reducing rating errors (Fay & Latham, 1982). Training in observation skills has been found to increase accuracy (Hedge & Kavanaugh, 1988). Other research has confirmed that student ratings are stable (i.e., consistent) over time (Hanges, Schnieder, & Niles, 1990). Simply put, “Will a student observe and record the same instructor behavior in the same way from one time to another and across different instructors?” Because consistency of ratings is a necessary (though not sufficient) step to ensure accuracy (Schmidt, 1990), additional training efforts to improve accuracy must be undertaken (i.e., Goals 1, 3, and 4). The training program for Goal 2 first revisits the questionnaire data that



were collected prior to Goal 1 training. Then, consistency of evaluations is taught using video Segments 3, 4, and 5.

Training Program

Goal 2 training begins by providing to students written learning objectives (see Table 3) and a handout with definitions of rater errors (e.g., Table 1). Use the results from the questionnaire about biases and attitudes to lead into a lecture about fairness of evaluations. Students should be told that the ultimate goal is to have accurate data about an instructor's performance and that any factors that reduce accuracy must be addressed. The training program begins with a short lecture about how biases and attitudes can result in reduced accuracy. Then, students have an opportunity to practice providing consistent evaluations and reducing biases in the evaluation process.

Table 3
Recommendations for Goal 2: Providing Fair Evaluations

Learning Objectives:

To consistently observe/record an instructor's performance
To consistently observe/record behaviors from one instructor to another
To demonstrate more complex processing of instructor performance data
To learn about the role of biases and attitudes in the evaluation process

Workshop Format:

Demonstration, Discussion, Practice, Feedback, Discussion

Video Segment 3:

Instructor Behaviors at Different Points in Classroom Lecture:
Learning to Consistently Observe and Record

Video Segment 4:

Several Instructors' Behaviors at Different Points in Classroom Lecture:
Learning to Consistently Observe and Record

Video Segment 5:

Several Instructors' Behaviors Varied by Level of Performance Across
Dimensions: Learning to Consistently Observe and Record

Format:

Lecture on Issues Tangential to the Evaluation Process



Pre- and Post-Test:

Change in Knowledge about Attitudes; Change in Attitudes

Suggestions:

Use video segments that reflect actual evaluation settings

Suggest students take notes during semester on observed instructor behaviors

Biases/attitudes. The lecture/discussion should address how each bias might be manifested. Students' early attitudes toward the course and instructor (Sauber & Ludlow, 1988), course workload (Greenwald & Gilmore, 1977b), and attitudes toward grades (Vasta & Sarmiento, 1970) have been related to student ratings at the end of a course. A student might rate an instructor high if he enjoys sports, as she does (similar-to-me effect). Or, a student rates all instructors high because he believes that the university would not hire someone who could not teach (leniency). Students should be warned about contrast effects as they undertake the task of observing and rating several instructors.

Two types of contrast effects are (1) knowledge about an instructor's previous performance (Gaugler & Rudolph, 1992) and (2) comparisons to instructors previously rated (Murphy & Balzer, 1986). Students may be particularly susceptible to the latter. Alert students to the contrast effect and tell them that an opportunity to discuss contrasts will be provided after they have completed Segments 4 and 5.

Goal 2 video segments follow the same approach as for Goal 1: demonstration, discussion, practice, feedback, and discussion. This short lecture is designed to address unintentional biases--biases that the student is not aware he or she is exhibiting. A more problematic issue is a student's intentional shifting of evaluation ratings. The use of student evaluation teams may help control for intentional rating inflation or deflation. Student evaluation teams are discussed further in the section entitled, "Administration Decisions."

Consistency across time. The trainer explains the concept of reliability and why reliability is a necessary condition for the valid use of student evaluations. A videotaped classroom lecture shows an instructor exhibiting the same behavior at one or more different times (Segment 3). The trainer leads a discussion about the similarity of behaviors from the beginning to the end of a classroom lecture and suggests how instructors might exhibit similar behaviors at different times during a semester. Students practice with another video segment, after which the trainer provides feedback on accuracy of student observations and recording. Students discuss their observations before moving to the next segment.

Consistency across instructors. The demonstration for Segment 4 shows different instructors exhibiting similar behaviors across all dimensions. Instructors might be teaching the same or different course material but the behaviors they exhibit are the similar, allowing for individual differences in style and approaches to teaching. The students practice on another video segment, receive feedback, and discuss accurate placement of several instructors' behaviors in dimensions.



More complex processing of performance information occurs in Segment 5 where different performance levels for several instructors are given. The demonstration and practice portions of the video should be prepared in a way that elicits attitudinal and contrast effects and most closely resembles what students might expect in the actual evaluation setting. After the practice segment is completed, discuss the impact of student attitudes, biases, and contrast effects on student observational and recording accuracy. Provide additional time to answer student questions about the questionnaire and what to expect in the actual evaluation setting.

Reassure students that the normal evaluation setting should allow for some control of a contrast effect. A delay between observing and rating has been shown to minimize contrast effects (Murphy, Balzer, Lockhart, & Eisenman, 1985). Remind them that the type of training received in Goal 1 (i.e., frame-of-reference) may also reduce contrast effects by providing prototypes for performance effectiveness (Murphy & Balzer, 1986). A more challenging issue may be training students to ignore an instructor's personality (Clayson & Haley, 1990). The design team should follow research recommendations for reducing the impact of student attitudes toward the instructor. For example, the evaluation system should be performance-based (see Tuckman, 1995, for a review). Behavioral examples (discussed under Goal 1) will focus student attention away from personalities and traits (Weirsmas et al., 1995). And, students should be encouraged to take notes on instructor behaviors throughout the semester. DeNisi and his colleagues found that diaries increase recall and rating accuracy and control for errors (e.g., halo, leniency) (DeNisi, Robbins, & Cafferty, 1989; DeNisi & Peters, 1992). Students would not need to keep extensive diaries but could use a behavioral checklist to note observed behaviors. In the evaluation setting, remind students to base their ratings on observed variability across instructors (Stamoulis & Hauenstein, 1993). Finally, students are reminded to recall this training material in subsequent Goals and in the actual evaluation setting.

Goal 3: Understanding the Broader Context of Instructor Evaluations

The purpose of Goal 3 is to increase student knowledge about the context in which their input is solicited. Several reasons support sharing information about the context. First, instructor behaviors and student evaluations don't occur in a vacuum. Training should not focus, therefore, only on how to reduce rater errors. Second, trainees are more likely to retain and recall information if they understand the broader context (Gagne & Briggs, 1979). Third, satisfaction with the performance evaluation system improves with additional knowledge (Giles & Mossholder, 1990). Fourth, more information results in better reliability across sources of data (Williams & Levy, 1992). Finally, credibility of the system may be enhanced (Nyirenda, 1994), resulting in students taking the process more seriously. (Also see Cardy & Dobbins, 1994, for a review of formats, cognitive schemata, and purposes of performance appraisal.)

The broader context includes the purpose of performance evaluation, the types of performance data collected, the sources of performance data, and how information is used (refer to Table 4). One might reasonably expect that broad university policies and procedures for use of student evaluations are established in a faculty handbook. College or department by-laws may guide procedures for collecting and using student evaluation data. At a minimum, students should know where they could access these policies and procedures.



Table 4
Recommendations for Goal 3:
Understanding the Broader Context of Instructor Evaluations

Learning Objects:

To know the purpose of evaluations

To know the sources of instructor evaluation information

To understand the full process of evaluation

Format:

Lecture on Broader Context

Suggestions:

Tell students how to find university policies on use of instructor evaluations

Provide for a transition to Goal 4 and to the actual evaluation setting

Training Program

Learning objectives are shared with the students. Students will have become familiar with evaluation criteria (Goal 1), but they may not know how the criteria were developed. Telling students if the evaluation criteria were developed internally and if students were involved may lend credibility to the process.

Students should also know the purpose of evaluations. The difficulty is that knowledge of the purpose of the rating can influence accuracy. When the purpose of evaluation is personnel decision making, ratings tend to be more lenient (Harris, Smith, & Champagne, 1995) and susceptible to bias (i.e., a popularity contest) (Tuckman, 1995). A developmental purpose results in increased feedback and [instructor] satisfaction with feedback (Tharenou, 1995). However, without knowing why performance evaluations are done, students may not take them seriously.

Students should understand that their evaluations are only one source of information on an instructor's teaching performance. Other sources include peer assessments and self-assessment. Peers are considered more appropriate sources of evaluation for some criteria, such as course planning/preparation and keeping up with teaching-related professional fields. Peers also provide input on delivery of instruction via classroom visits. Though peer reviews have been criticized on a number of points, they are reliable and valid indicators of performance (Latham & Wexley, 1994). Self-assessment includes an examination of strengths and weaknesses resulting in a developmental plan for improving teaching effectiveness. To reduce the possibility of a leniency effect, Fahr, Werbel, and Bedeian (1988) suggest that self-appraisals be supported by objective data. Instructors might prepare a portfolio of classroom-related materials and assessment methods (e.g., syllabi, exams, student projects). Participation through self-assessment may increase the instructor's acceptance of the process and accountability for improving (see an expanded discussion of accountability in the section on Administration Decisions).



Finally, students may be informed about how student evaluations are used. Though specific student feedback may be looked at more closely at lower levels in the university (i.e., department), if students understand the potential impact of their evaluations on the short- and long-term development plans for the instructor (and hence achievement of department and university vision), they may take them more seriously. Further, upward feedback [from students] can be used to reinforce team values (Auteri, 1994).

Summary of Goals 1, 2, and 3

Training in the first three Goals should prepare students adequately to conduct evaluations during their first year at the university. Encourage students to make a transition into the actual setting by allowing student trainees to visit actual classes early in the semester to practice their newly-acquired rating skills. The typical evaluation process and setting should be explained (e.g., evaluation surveys are administered by the testing center in the actual classroom) as well as unique practices (e.g., criteria for evaluation vary by department). Finally, students should be told that the retraining they will receive throughout their college career will provide an opportunity to discuss their experiences and address particularly challenging evaluation settings (e.g., pedagogical approaches such as cooperative learning groups or lecture may make it challenging to compare instructors).

Goal 4: Preparing for the Evaluation Process

Goal 4 is created to retrain students periodically in important aspects of the evaluation process. While the training that students received in Goals 1, 2, and 3 has a priming effect that encourages recall at the appropriate times, additional cues need to be provided to ensure that students remember their training. Simple reminders in the evaluation instructions will cue recall (Anderson, 1985). For example, “using the dimensions for evaluating instructor performance, complete the evaluation form, considering the instructor behaviors that you observed during the semester. Recognize the types of biases that can interfere with evaluations and avoid them.” However, Goal 4 recognizes that the effects of rater training decrease after 6-12 months (Ivancevich, 1979). Martin and Bartol (1986) recommended annual retraining. Besides cueing, retraining is also an important opportunity to share changes that may have occurred in dimensions (Goal 1) or administrative procedures (Goal 3).

Table 5
Recommendations for Goal 4: Preparing for the Evaluation Process

Learning Objectives:

To recall dimensions, terms, and behaviors (Goal 1)

To recall the importance of providing fair and accurate evaluations (Goal 2)

To recall the context within which evaluations are conducted (Goal 3)

Format:

Discussion; Question and Answer



Suggestions:

Provide visual cues to enhance recall; provide cues in evaluation instructions
Allow time for small group discussion and sharing real evaluation experiences

Training Program

First, share the learning objectives with students (see Table 5). Then, begin the training. Since Goal 4 involves retraining, visual cues may be sufficient to induce students to recall material from Goals 1, 2, and 3. Overhead transparencies could summarize the important points. Goal 4 is an appropriate time to use student groups for discussions (e.g., small groups of 4-5 students, with 5-6 groups in a training session). Training sessions lasting 1.5 hours may be sufficient (for Goal 4).

Annual retraining need not be as extensive as training in the previous three Goals, but it does require a different kind of information processing. Students may have questions about their training in Goals 1, 2, and 3 and based on their first-year experience with conducting evaluations. Be sure to allow for sufficient group discussion time to generate student questions and to allow students to share real situational experiences. The role of the trainer is to provide clarification answers and to help students deal with actual rating situations.

Retraining of returning students can be conducted on the same day that new students receive training (i.e., new student orientation). Expect some resistance from returning students who are now required to attend an additional orientation session. Resistance might be lessened by holding retraining sessions in residence halls or at the unit level (i.e., college or department). Resistance might come from those who are responsible for coordinating the training. For example, documenting whether or not students attended the required training for all four Goals is time-consuming. A more appealing approach may be to create mechanisms by which students and instructors are motivated and accountable for involvement in the development and implementation phases of this training program.

Administrative Issues

Several administrative issues need to be addressed to accomplish the suggested training. Mechanisms must be put into place to show commitment to the training. Design and implementation teams must be formed to identify (or develop) criteria, to design the training, and to coordinate and implement the training (refer to Table 6). Some of these issues are elaborated below.

Table 6
Administrative Issues in Design and Implementation
of the Training Program

Cost/Benefit Analysis:



Number of New Students (Goals 1, 2, and 3)
Number of Returning Students (Goal 4)
Determine Training Method - Workshop, Lecture, Discussion, Q/A
Determine How to Motivate Student/Instructor Participation
Measure Increased Student Accuracy, Improved Decision Making, and Morale

Forming a Design Team:

Identification (or further development) of Criteria/Behaviors
Creation of Training Modules for Goals 1, 2, 3, 4
Creation of Videos
Creation of Lectures
Creation of Training Module to Train the Trainers

Training the Trainers:

Select Internal Trainers: Instructors versus Staff
Determine When to Train

Coordinating Training:

Determine When to Offer Goal 1, 2, 3 Training to New Students
Determine When to Offer Goal 4 Training to Returning Students
Find Suitable Location Based on Training Method

Administration Decisions

Universities demonstrate their commitment to this process in a number of ways, some of which are resource-related. By including accountability issues in vision and mission statements, the university signals the value of time spent on developing policies and procedures that reinforce accountability at all levels in the university. A broad discussion of accountability is beyond the scope of this article; however, student accountability for conscientiousness in the evaluation process can be encouraged by administration.

In the context of ensuring accuracy of student evaluations, instructors and students should be involved in the design, development, and implementation phases of the training program. Instructor participation reinforces acceptance of the system. Student participation reinforces accountability for accuracy. If students are held accountable, they may, in part, be motivated to attend training and retraining sessions. The essence of the discussion that follows would be communicated to students when they are told that they must attend training and retraining sessions.

Research has supported that raters are more vigilant, are more accurate, and take the process more seriously when they are held accountable (Hauenstein, 1992; Mero & Motowidlo, 1995). London, Smither, and Adsit (1997) suggested that anonymity can be maintained and accountability can be moderately increased by using a facilitator to encourage discussions



between the rater and the person being rated. In the student/instructor evaluation setting, universities are encouraged to adopt a student evaluation team approach for providing feedback to instructors during the semester. Students bring instructor-, course-, or classroom-related issues to the student evaluation team for discussion. The team, in turn, presents the issues to the instructor at pre-set times throughout the semester. No one is identified directly, but the student evaluation team does require individual students to be clear and accurate with their behavioral examples. Intentional distortion of ratings be minimized, and if the instructor uses this feedback to improve the course, students are reinforced to provide course-appropriate evaluations. This form of student evaluation and feedback may or may not replace the more traditional end-of-course evaluation, but if handled correctly, can contribute to accountability in the evaluation process.

Method/Length of Training Sessions

As suggested earlier, the number of students that need training will determine the cost associated with implementing this training. Workshop methods are the most time-consuming; lectures are the easiest way to share information with large numbers of students. Discussion groups can be incorporated into both workshops and lectures. An important aspect in training for Goals 1, 2, and 3 is the opportunity for students to receive feedback on their responses. If forms of feedback can be provided in a lecture and discussion format, some improvement in accuracy will occur (Stamoulis & Hauenstein, 1993).

In research, the length of training has varied from as little as one hour (Bernardin & Walter, 1977) to as many as fifteen hours (Martin & Bartol, 1986). Hedge and Kavanaugh (1988) concluded that two hours of training were not enough. Universities will need to determine a break-even point that establishes how much training is required to generate improvements in reliability and accuracy of student ratings and faculty/instructor morale (also see Cost-Benefit Analysis, below).

Internal/External Trainers

Internal trainers are recommended. Universities have a wealth of expertise at their fingertips. Advantages of internal trainers include knowledge of the university, lower costs, instructor buy-in (through participation/involvement), and improved transfer of training because follow-up questions can be directed to an on-site trainer. Instructors or staff can provide the training. In either case, trainers should be trained in the Goals and methods for achieving the goals. Though each of these phases could be completed at the unit level (i.e., college or department), university-level planning and design are encouraged to ensure consistency.

Design Team

In the simplest sense, a design team would represent academic disciplines and include persons with expertise in designing performance evaluation systems. Student representatives (e.g., from student governance groups) would be welcome additions to this team. Instructors should be involved in the identifying and specifying of performance dimensions and measurement scale anchors to promote buy-in of the final product.



The design team responsibilities begin with identifying (or further developing) criteria for evaluation. Since Smith and Kendall's (1963) introduction of behavioral scaling methods, several behavioral-based approaches have been proposed over the years, including behaviorally-anchored rating scales, behavioral expectation scales, behavioral observation scales, and behavioral summary scales (see summaries by Cardy & Dobbins, 1994, and Peters & DeNisi, 1990). The design team creates the training to accomplish the Goals. Borman (1977) and Latham, Wexley, and Pursell (1975) provide instructions on creating videos and guidelines for workshop development. The design team is also responsible for determining when and where to hold the training sessions. Finally, the design team provides support to the implementation team to coordinate and schedule the actual training activities.

An overall summary of the process for designing and implementing the training program is provided in a checklist format in Table 7.

Table 7
A Design and Implementation Check List

Design

- ☐ A design team is formed including instructors, students, administration, and experts
- ☐ A plan for evaluating the effectiveness of the training program is developed
- ☐ Criteria and standards for performance are determined
- ☐ The total number of students to be trained and retrained are identified
- ☐ Trainers are identified and trained
- ☐ The methodology for conducting training sessions is determined (pedagogy/length)
- ☐ Training programs are designed and materials/supplies are developed/obtained
- ☐ Rooms and trainers are schedule

Implementation

- ☐ Students complete short questionnaires on attitudes (held for Goal 2 discussions)
- ☐ Goal 1 learning objectives are provided; appropriate pre- and post-tests are given
- ☐ Goal 2 learning objectives are provided; appropriate pre- and post-tests are given
- ☐ Goal 3 learning objectives are provided; appropriate pre- and post-tests are given
- ☐ Goal 4 learning objectives are provided; appropriate pre- and post-tests are given
- ☐ Cost-benefit analysis is conducted
- ☐ Modifications to training programs and processes are made

Cost-Benefit Analysis

The usefulness of a new training program, such as the one recommended, must be evaluated to determine if improvements outweigh the expenses associated with the design and implementation of the training program. The cost-benefit analysis considers (a) the relative importance of teaching in the broader university setting, and (b) the measurable outcomes and expenses associated with the actual training program.



The overall usefulness of the training program for student evaluators will be a function of the relative importance of teaching in the university setting, as well as the relative importance of teaching in the instructor's overall responsibilities. First, teaching universities may find more benefits to the training program than universities that emphasize research. Some universities place a greater relative emphasis on teaching regardless of research and/or service activities. Second, the training program would have a greater usefulness to an instructor whose sole job is to teach. As other responsibilities increase (i.e., research, internal service, external service) and teaching responsibilities decrease (e.g., a 6-hour teaching load compared to a 12-hour teaching load), the overall usefulness of the training program goes down. The following discussion highlights the multiple aspects of benefits and costs.

Benefits

Benefits are tangible and intangible. Measurable benefits include improved student accuracy leading to improved accuracy of personnel decisions and increased student knowledge. As suggested by previous research, more accurate evaluations will result from training. More accurate input into personnel decisions will increase the likelihood that correct personnel decisions (e.g., retention, termination, need for instructor development) will be made. Intangible benefits include more positive perceptions of fairness, improved instructor morale, and student acquisition of life-long skills. Perceptions of fairness and morale are believed to lead to improved productivity. [A discussion about measuring the impact of attitude change on productivity is beyond the scope of this article. Fortunately, other sources exist that explain how to conduct a study (see Cascio, 1991, pages 130-148.) Finally, the life-long knowledge about performance appraisal systems and the acquisition of rating skills that students gain may be highly valued by organizations concerned about fairness, legal compliance, and ethical issues.

Costs

Administration will be concerned with the fixed and variable costs for the training of student evaluators. The identification (or development) of criteria, training design, and training the trainer activities represent fixed costs. Implementation of training is a variable cost based on length/number of training sessions, how many students will be trained and retrained, whether internal (e.g., instructors, staff) or external trainers will conduct the training, and other administrative expenses (e.g., scheduling training sessions, assigning students to training sessions).

Costs are mainly tangible, though opportunity costs (Cascio, 1991) must not be overlooked. Tangible costs include faculty and staff time, supplies, and other overhead expenses (e.g., heat, light). Faculty and staff involvement in the design and delivery of the training program can be measured as either lost productivity (i.e., being drawn away from other activities) or as actual expenditures (i.e., receiving extra compensation for their involvement). The design and development of the training program for student evaluators is a one-time cost, though periodic review and updating may be necessary. Additional expenses accrue if external consultants are hired to provide special expertise. The implementation of the training program will include costs associated with printing and duplication of presentation materials, packets/notebooks for students, and general maintenance of training facilities.



Intangible costs may include the probability of a grievance or a lawsuit if the university bases its decisions on data that are not reliable and valid. Because personnel decisions within the university setting are confidential, it will be difficult to calculate this probability. Further, while lawsuits tend to be more public, the extent to which unreliable and invalid student evaluations were the basis for a personnel decision may be difficult to determine. Finally, the university must consider the cost of not training student evaluators against the cost of not spending monies on other programs (i.e., lost opportunities).

Summary

The costs of a training program for student evaluators will have to be weighed against the benefits received. This analysis is specific to each university for administration must consider the relative emphasis of teaching to the university (as a whole) and to the faculty/instructors (as individuals). Flexibility exists for the design and implementation phases as universities find a break-even point where improvements begin to outweigh costs. While the training program presented here suggests multiple training sessions focusing on four Goals, implementation efficiencies exist (see Administrative Issues, above).

Conclusion

This article suggests that previous research findings on how to improve performance evaluation ratings can be applied to student evaluations of instructor performance. The performance evaluation research has focused on improving accuracy by understanding the rater/s cognitive processes. Research on student evaluations identified the multiple dimensions of student ratings and contaminants in the evaluation process. A unified training program for student evaluators may result in more reliable and accurate evaluations and increased acceptance of the process by instructors.

By providing the training when students first enter the university, benefits will accrue within and outside the educational setting. This article argues that the need for accuracy of student evaluations is not only intuitive, but also paramount--accurate input is a necessary condition for valid use of the information. Each university will have to evaluate the costs and benefits of designing and implementing a training program for student evaluators of instructors. Hopefully, the ideas presented will spark an interest in exploring the benefits of developing and implementing training for student evaluators.



Appendix

Based on principles of learning, events of instruction include (Gagne & Briggs, 1979):

- Event 1: Gain Attention
- Event 2: Inform Trainee of Learning Objective
- Event 3: Stimulate Recall of Prerequisite Learning
- Event 4: Present Training Material
- Event 5: Provide Learning Guidance
- Event 6: Elicit Trainee Performance
- Event 7: Provide Feedback to Trainee
- Event 8: Assess Trainee Performance
- Event 9: Enhance Retention of Learning and Transfer of Training

A general introduction, during which no important information is shared, serves well as a transition (Event 1). Introduce the trainer, provide a brief background about the development of the training program, and explain what students will learn during the training. Provide written learning objectives for all Goals (Event 2). Suggest that students' job experience may have already familiarized them with evaluation concepts (Event 3). Goals 1 and 2 use 3-5 minute video segments (Event 4); Goals 3 and 4 use lecture and discussion. The design team may have to combine Segments 1-5 if resource constraints exist. The videos should be structured to be as similar as possible to those that the student will encounter in the classroom (Event 9); tell students that they are expected to recall training in subsequent Goals as well as in the actual evaluation setting. Require students to practice the above steps (Event 6). Students have better recall if they receive feedback and are actively involved in the training process. Feedback from the trainer includes acknowledging accuracy and providing ways for students to improve skills (Event 7). Combine events where possible. For example, learning guidance (Event 5) might include immediate feedback. Finally, a well-designed research study (Event 8) will determine that the training has accomplished the Goals.



References

- Anderson, J. R. (1985). Cognitive psychology and its implications (2nd ed). New York: W. H. Freeman and Company.
- Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels of processing theory and social facilitation theory perspectives. Journal of Applied Psychology, 72(4), 567-572.
- Auteri, E. (1994). Upward feedback leads to culture change. HR Magazine, 39(6), 78-84.
- Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. Journal of Educational Psychology, 72, 593-604.
- Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62(1), 64-69.
- Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential biases. Journal of Applied Psychology, 74, 478-494.
- Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Processes, 20, 233-252.
- Borman, W. C., & Dunnette, M. D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.
- Cardy, R. L., & Dobbins, G. H. (1994). Performance appraisal: Alternative perspectives. Cincinnati, OH: South-Western Publishing Co.
- Clayson, D. E., & Haley, D. A. (Fall, 1990). Student evaluations in marketing: What is actually being measured? Journal of Marketing Education, 9-17.
- Cronbach, L. J. (1955). Processes affecting scores on "understanding others" and "assumed similarity." Psychological Bulletin, 52, 177-193.
- Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. Journal of Applied Psychology, 80(1), 158-167.
- DeNisi, A. S., & Peters, L. J. (May, 1992). Diary keeping and the organization of information in memory: A field experiment. Poster session presented at the Seventh Annual Conference of the Society for Industrial and Organizational Psychology, Montreal, Quebec.



- DeNisi, A. S., Robbins, T. L., & Cafferty, T. P. (1989). Organization of information used for performance appraisal: Role of diary-keeping. Journal of Applied Psychology, 74, 124-129.
- DeNisi, A. S., & Williams, K. J. (1988). Cognitive approaches to performance appraisal. Research in Personnel and Human Resource Management, 6, 109-155.
- Fahr, J. L., Werbel, J. D., & Bedeian, A. G. (1988). An empirical investigation of self appraisal-based performance evaluation. Personnel Psychology, 41, 141-156.
- Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. Personnel Psychology, 35, 105-116.
- Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. Research in Higher Education, 28, 291-344.
- Gagne, R. M., & Briggs, L. J. (1979). Principles of instructional design (2nd ed.). New York: Holt, Rinehart and Winston.
- Gaugler, B. B., & Rudolph, A. S. (1992). The influence of assessee performance variations on assessor's judgment. Personnel Psychology, 45, 77-98.
- Giles, W. F., & Mossholder, K. W. (1990). Employee reactions to contextual and session components of performance appraisal. Journal of Applied Psychology, 75(4), 371-377.
- Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings. American Psychologist, 52, 1182-1186.
- Greenwald, A. G., & Gilmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. American Psychologist, 52, 1209-1217.
- Greenwald, A. G., & Gilmore, G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. Journal of Educational Psychology, 89 (4), 743-751.
- Hanges, P. J., Schneider, B., & Niles, K. (1990). Stability of performance: An interactionist perspective. Journal of Applied Psychology, 75, 658-667.
- Harris, C. (1988). A comparison of employee attitudes toward two performance appraisal systems. Public Personnel Management, 17, 443-456.
- Harris, M. M., Smith, D. E., & Champagne, D. (1995). A field study of performance appraisal purpose: Research versus administrative-based ratings. Personnel Psychology, 48(1), 151-160.



- Hartel, C. E. (1993). Rating format research revisited: Format effectiveness and acceptability depend on rater characteristics. Journal of Applied Psychology, 78(2), 212-217.
- Hauenstein, N. M. A. (1992). An information-processing approach to leniency in performance judgments. Journal of Applied Psychology, 77, 485-493.
- Hedge, J. W., & Kavanaugh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraisal training. Journal of Applied Psychology, 73(1), 68-73.
- Heneman, R. L., Wexley, K. N., & Moore, M. L. (1987). Performance rating accuracy: A critical review. Journal of Business Research, 15(5), 431-448.
- Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. Journal of Applied Psychology, 64, 502-508.
- Kinicki, A. J., Hom, P. W., Trost, M. R., & Wade, K. J. (1995). Effects of category prototypes on performance-rating accuracy. Journal of Applied Psychology, 80(3), 354-370.
- Latham, G. P., & Wexley, K. N. (1994). Increasing productivity through performance appraisal (2nd ed.). Reading, MA: Addison-Wesley.
- Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize ratings errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.
- London, M., Smither, J. W., & Adsit, D. J. (1997). Accountability: The Achilles' heel of multisource feedback. Group & Organization Management, 22(2), 162-184.
- Marks, R. (2000). Determinants of Student Evaluations of Global Measures of Instructor and Course Value, Journal of Marketing Education, 22 (2), 108-199.
- Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76(5), 707-754.
- Martin, D. C., & Bartol, K. M. (1986). Training the raters: A key to effective performance appraisal. Public Personnel Management, 15(2), 101-109.
- Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. Journal of Applied Psychology, 80, 517-524.



- Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. Journal of Applied Psychology, 71, 39-44.
- Murphy, K. R., Balzer, W. K., Lockhart, M., & Eisenman, E. (1985). Relationship between observation accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.
- Nyirenda, S. (1994). Assessing highly accomplished teaching: Developing a metaevaluation criteria framework for performance-assessment systems for national certification of teachers. Journal of Personnel Evaluation in Education, 8, 313-327.
- Park, O. S., & Sims, H. P. (1989). Beyond cognition in leadership: Prosocial behavior and affect in managerial judgment. Paper presented at the Academy of Management Meeting, Washington, D. C.
- Peters, L. H., & DeNisi, A. S. (1990). An information processing role for appraisal purpose and job type in the development of appraisal systems. Journal of Management Issues, 2(2), 160-175.
- Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., & McCormick, S. (1995). Direct, indirect, and controlled observation and rating accuracy. Journal of Applied Psychology, 80(6), 664-670.
- Sauber, M. H., & Ludlow, R. R. (1988). Student evaluations stability in marketing: The importance of early class meetings. The Journal of Midwest Marketing, 3, 41-49.
- Schmidt, J. J. (1990). Critical issues for school counselor performance appraisal and supervision. School Counselor, 38(2), 86-94.
- Schneider, B., Hanges, P. J., Goldstein, H. W., & Braverman, E. P. (1994). Do customer service perceptions generalize? The case of student and chair ratings of faculty effectiveness. Journal of Applied Psychology, 79(5), 685-691.
- Smith, D. E. (1986). Training programs for performance appraisal: A review. Academy of Management Review, 11(1), 22-40.
- Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.
- Stamoulis, D. T., & Hauenstein, N. M. A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. Journal of Applied Psychology, 78(6), 994-1003.



- Tharenou, P. (1995). The impact of a developmental performance appraisal program on employee perceptions in an Australian federal agency. Group & Organization Management, 20(3), 245-261.
- Tuckman, B. W. (Winter, 1995). Assessing effective teaching. Peabody Journal of Education, 70(2), 127-138.
- Vasta, R., & Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance. Journal of Educational Psychology, 71, 207-221.
- Weirsmas, U. J., Vandenberg, P. T., & Latham, G. P. (1995). Dutch reactions to behavioral observation, behavioral expectation, and trait scales. Group and Organizational Management, 20(3), 297-309.
- Williams, J. R., & Levy, P. E. (1992). The effects of perceived system knowledge on the agreement between self rating and supervisor rating. Personnel Psychology, 45, 835-847.
- Woehr, D. J. (1992). Performance dimension accessibility: Implications for rating accuracy. Journal of Organizational Behavior, 13(4), 357-367.