AN OVERVIEW OF TEACHER EVALUATION*

Cathy Barrette, Elaine Morton, Anjel Tozcu
University of Arizona

Although teacher evaluation is a crucial aspect of the educational process, there are many questions as to the best way to carry it out. Among the relevant issues that we discuss in this paper are the purposes, content, development, and administration and scoring of teacher evaluations. Important decisions are made on the basis of information obtained during the teacher evaluation process; hence it is of paramount importance that the assessment methods and instruments be both valid and reliable. In an effort to shed light on this vital area of pedagogy and program administration, we provide an overview of the professional literature regarding effective teaching, and critically evaluate various current methods and instruments. Finally, we recommend strategies for overcoming the many difficulties encountered by those implementing teacher evaluation processes.

INTRODUCTION

Teacher evaluation is a vital part of the educational process, yet there is no consensus on the best way to carry it out. Most education professionals would agree, however, that since many important decisions are made on the basis of information gathered in the evaluation process, it is crucial that the instruments used be both valid and reliable. The goal of this paper is to present a historical perspective of teacher evaluation, and a discussion of the various evaluation procedures and criteria currently being used, in order to make some recommendations on how to maximize the validity and reliability of teacher evaluations.

BACKGROUND

Informal student evaluation of teachers began as early as the fifteenth century, when students at the University of Bologna paid instructors according to their teaching abilities (Annadale, 1974, p. 11). Haefele (1993) cites research on teacher evaluation in the USA dating back to at least 1915. Marsh and Bailey (1993) state that "the literature on students' evaluations of teaching effectiveness (SETE) consists of thousands of studies and dates back to the 1920s and earlier" (p. 1). Annadale (1974) mentions that "Harvard's Confidential Guide," an informal student evaluation of their teachers, was distributed as early as 1924.

In studies in the late 1920s, students and expert evaluators were asked to describe teachers they considered to be effective, and to rate characteristics of good teachers. In the 1930s, scales were devised for the evaluation of teachers; these scales were based on qualities believed to be important in teaching (Nerenz & Knop, 1982, p. 243).

As a preliminary step in developing a system of teacher evaluation, Troyer and Pace (1944) report that Columbia University formed a committee to formulate criteria that would later serve as the basis for teacher evaluations. The criteria consisted of a set of principles and objectives for teachers. For example, one of the principles was that both peer and student evaluations should be included in the evaluation process; in general, the objectives promoted personal as well as professional growth of teachers.

Prior to the 1960s, emphasis was placed on the personal characteristics of good teachers, such as age, sex and race. However, in the 1960s, the focus began to shift to teachers' behavior in the classroom because teacher behaviors were believed to have a direct effect on student achievement. Thus, both the teaching process and the product of teaching, in terms of student outcomes, became prominent issues in teaching evaluation (Nerenz and Knop, 1982).

By the end of the 1960s, evaluators became concerned with making the evaluation process more objective (Delamere, 1986), as seen by proposals for using pre-specified categories as the basis for observations of teachers (Flanders, 1970). A large number of studies on teacher

* We would like to express our appreciation to Dr. Renate Schulz for her helpful comments on an earlier version of this paper.

evaluation concerned with validity and reliability were conducted in the 1970s. Evidence for the concern with reliability and validity includes the 31 studies discussed by Benton (1982), and additionally, Aleamoni (1987b) cites 74 studies; of these, 48 were published in the 1970s, and more than half (29) of these dealt with the issues of reliability and validity. Concern with these issues is still in evidence today in such articles as Akpe and Igwe (1992), Haefele (1993), MacGinitie (1993), Marsh and Bailey (1993) and Wennerstrom and Heiser (1992).

TEACHER EFFECTIVENESS AS A CENTRAL CONCEPT IN TEACHER EVALUATION

The first step in developing a valid instrument requires defining the construct to be measured, which in this case is teacher effectiveness. Teacher effectiveness is an elusive construct, and therefore difficult to define, as will be seen in the various definitions given below. In fact, according to Feldvebel (1980), for most evaluation systems that are not professionally developed, no overt attempt is made to provide an explicit definition of teacher effectiveness. However, he reviews four kinds of criteria that have previously been associated with teacher effectiveness.

In the first group of criteria, certain characteristics of the teachers, such as social status, age, sex, race, and education, were formerly perceived to be predictive of teacher effectiveness, and were therefore considered to be valid criteria for determining teacher effectiveness. However, such traits are situation specific, and not valid predictors of teacher effectiveness (Feldvebel, 1980, pp. 416-417). In fact, in the case of an ESL classroom, some of these factors have been shown to bias student evaluation of teachers, "due most significantly to level, course type, ethnic background, and attitude" (Wennerstrom & Heiser, 1992, p. 283-284). Thus, these traits cannot be considered a valid way of identifying effective teachers.

Another perspective, also dealing with teacher characteristics, establishes a set of criteria for effective teaching based on of a list of competencies. Pennington and Young (1989) list competencies (based on Delamere, 1986, and TESOL literature) that are generally believed to be associated with effective ESL teaching. The main categories are knowledge; skills; and, attitudes (p. 625). However, to know if these competencies underlie the construct of teacher effectiveness, it would have to be determined that a teacher who has these competencies is actually effective in the classroom. Similarly, Brown (1994) presents a checklist of Good Language Teaching Characteristics, based on TESOL's certification guidelines and unpublished sources. The main categories are technical knowledge; pedagogical skill; interpersonal skills; and, personal qualities (pp. 429-430). Brown recommends that these characteristics be used by language teachers as a self-check to find areas for continued professional growth and goal-setting.

In the second group of criteria, the focus is on teacher behaviors because of their perceived influence on learning outcomes. For example, descriptive instruments such as Flanders' interaction models (Flanders, 1970) enable the evaluator to categorize types of classroom behaviors in order to objectively identify interaction patterns. Recognition of these patterns by the teacher can increase teacher self-awareness and promote professional growth. Difficulties arise with construct validity, however, because the relationship between teacher effectiveness and classroom interaction is unclear (Feldvebel, 1980, p. 417). Also, this purported relationship is based on the misconception that teaching done in one particular way will result in effective learning regardless of context (Nunan, 1989, p. 98), and for any student. In addition, due to the differing goals of various instructional activities, behavior patterns cannot be reliably compared from one activity or group of activities to another.

In the third criteria set, student achievement has been used as a measure of teacher effectiveness (Feldvebel, 1980, p. 418). However, this use is problematic because of the low construct validity; many factors in student achievement are beyond the teacher's control, and are therefore unrelated to the teacher's behavior. Furthermore, Pennington and Young (1989) state that "research on the reliability of student test scores as a measure of teaching effectiveness has consistently indicated reliability to be quite low" (p. 630). This use of measures of student achievement will be discussed more fully at a later point in this paper.

The last set of criteria includes contextual factors, such as location, available materials, student demographics, institutional framework, and community input. These can all work together to influence both teachers' actions and student behavior and learning outcomes. Unfortunately, few evaluation instruments take such contextual elements into account (Feldvebel, 1980, p. 418) even though they can frequently act as confounding variables in teacher evaluation (Tracey, 1978).

In determining the parameters they will use to define teacher effectiveness, programs may profitably refer to some of these criteria. However, each program must carefully define teacher effectiveness according to their particular context, and must then make sure that the criteria they decide upon are clearly written and available to all faculty and evaluators.

IMPORTANT FACTORS IN CONSTRUCTING A TEACHER EVALUATION INSTRUMENT

Based on the foregoing discussion of criteria previously used in defining teacher effectiveness, we will demonstrate the importance of establishing a definition of teacher effectiveness, and mention some factors shown to correlate with it. Furthermore, some factors important to consider in constructing a valid and reliable teacher evaluation system will be proposed. These include establishing the purpose for the evaluation; carrying out a job analysis to determine the appropriate aspects of teaching to be evaluated and deciding on the relative importance of each of these aspects; standardizing the evaluation procedure; and assuring the validity and reliability of the entire process.

The definition of teacher effectiveness is central to the process of teacher evaluation, yet our review of the literature indicates that there is no single, widely-accepted definition of this construct. For the purpose of demonstrating the importance of defining this construct, we will utilize some examples of actual definitions that have been used, but we do not necessarily advocate the use of these definitions.

In the days of Aristotle and Socrates, an effective teacher was one who could attract students; at the University of Paris, founded in the tenth century, the number of students who chose to study under a particular instructor was also seen as a measure of teacher effectiveness (Travers, 1981, p. 14). Thus, based on this definition, a teacher evaluation might well have consisted of counting the number of students a particular teacher was instructing. In contrast, modern conceptions of teaching recognize that it "is obviously a complex activity with many components working together to make a successful class" (Wennerstrom & Heiser, 1992, p. 272). The Michigan State University Student Instructional Rating System (SIRS), for example, includes five aspects of teacher effectiveness: "instructor involvement, student interest and performance, student-instructor interaction, course demands, and course organization" (p. 272). All of these factors must be measured in order to have a valid instrument for this definition. In fact, it is likely that the definition of teacher effectiveness will depend to a large extent on the particular context in which the teaching occurs.

As important as it is to define teacher effectiveness, it is equally important to establish the purposes for which the evaluation will be used because the validity of the evaluation instrument is a function of its purpose. Evaluations can have either a formative purpose, as when they are used to improve teaching quality by identifying strengths and weaknesses, or a summative one, as when they are used in making personnel decisions. If used for formative purposes, there should be provision in the evaluation procedure for discussion of the results of the evaluation so that teachers can make use of the feedback they receive to improve the quality of their teaching. On the other hand, if used for summative purposes, evaluators must pay close attention to whether or not the decisions that are based on the evaluations are legally defensible and ethically sound. For example, instruments designed for formative purposes, such as Carroll's (1981) and Christison and Bassano's (1984) self-evaluation models, may not be appropriate for summative purposes due to the subjectivity involved.

The purpose of a job analysis is to objectively identify the duties and activities that comprise a job (Haefele, 1993). A list of "key performance areas" (Tracey, 1978, p. 241) can be derived from the job analysis, although what MacGinitie (1993) terms the "measurement selection

problem" can cause difficulties at this stage (p. 558). This is the problem of deciding which aspects of performance are important enough to measure. Once this list of activities is established, it can serve as an objective measure of task completion, thereby facilitating comparison of several individuals who hold the same position. Since not all responsibilities are of equal importance, certain aspects of the job should be given more weight than others. For example, Magnusen (1987) reports on an evaluation system that includes sections on teaching, research, and service. Although in this example each section is equally weighted, research carries more weight than do teaching and service in some Research I institutions.

Standardizing the evaluation process is crucial if one hopes to achieve validity and reliability. The instruments used, the administration process, scoring procedures, and criteria for the interpretation and use of the evaluation must be standardized in order to permit fair and equitable rating of teachers. Aleamoni's (1978) Course/Instructor Evaluation Questionnaire (CIEQ) is an example of an evaluation instrument that was designed to minimize the possible problems arising from the use of student evaluation of teachers, and has been shown to be both valid and reliable.

Careful attention to the above-mentioned factors will contribute to the development of an evaluation process that meets the requirements for valididty and reliability. In summative uses, reliability and validity are especially crucial, since decisions made on the basis of the results of the evaluation must be both ethical and legally defensible. Based on a review of research, Haefele (1993) concludes that to be legally defensible, an evaluation must have written policies that are made available to all persons involved in the evaluation process; the administration and scoring must be standardized; a job analysis must be the basis of evaluation criteria as well as of the weight given to each criteria; and, the evaluations must be carried out on multiple occasions by more than one trained rater (p. 23).

CURRENT EVALUATION METHODS

Teacher Interviews

Teacher interviews consist of standardized rank-ordered questions asked by a supervisor in private. They can be used for employment selection, where predetermined questions are asked. If based on the program's teacher effectiveness criteria, this can be a valid part of pre-employment evaluation. However, the atmosphere may seem impersonal with verbal interaction limited to set questions, and it may be expensive to train interviewers. Interviews are also used to review performance following administrator observation so that the teacher and the administrator can collaboratively evaluate the performance and set goals. Interviews have the advantage of being confidential; however, they are time-consuming, and interviewers may be biased, although the use of multiple interviewers can offset this bias problem (Pennington & Young, 1989, pp. 621-623).

Competency Tests

Competency tests consist of a standardized test battery; these are mainly given as part of the teacher certification process, mainly for summative purposes. They measure communication skills, professional knowledge, content area knowledge, and knowledge of teaching methodology. The advantages of these tests are that they are standardized and objectively scored, legally defensible, economical to administer, and assure a minimum level of knowledge. However, they have low predictive validity for teaching effectiveness, and thus lack construct validity. Furthermore, they have been found to be unreliable for experienced teachers (Pennington & Young, 1989, pp. 624-626).

Student Achievement

Student achievement has been used for summative purposes as a measure of teacher effectiveness. Advantages are that it is objective, quantitative, economical, and has high face

validity. However, this method has low reliability as well as low construct validity. Since many factors that are not within the control of the teacher affect students' performance, student achievement does not equate with teacher effectiveness (Pennington & Young, 1989, pp. 630-631). Also, teachers judged on this basis may teach to the test, causing a negative washback effect, with less creative teaching and narrower course content resulting.

Class Observation by a Supervisor

Class observation by a supervisor is used both summatively and formatively, and consists of the supervisor rating teaching performance observed during a classroom visit. Usually there is a pre-observation meeting in which the observer gets information on the lesson to be observed, evaluation criteria are agreed upon, and procedures are discussed. Advantages are that the evaluation is based on relevant, observable criteria which often have been determined by faculty and administrators together; also, the teacher usually receives post-observation feedback. If observers have been properly trained, this can be a valid and reliable method (Pennington & Young, 1989, p. 634). However, in cases where evaluators are not trained, they can be judgmental and biased as well as unsystematic. Teachers may feel threatened by the presence of a supervisor in their classroom (Sheal, 1989). In addition, the evaluator may be unfamiliar with the content or language being taught, and contextual factors may be ignored (Pennington & Young, 1989, pp. 634-637). Lastly, since this method is very time-consuming, it is costly. (For further discussion of supervisor evaluation see Fanselow, 1988; Pope, 1990; Wood, 1992.

Peer Review

Peer review, commonly used at the university level, can be used both summatively and formatively, and usually involves pre- and post-observation conferencing. Advantages of this method are that it has the potential for establishing peer networks that encourage idea sharing (Barnett, 1983); evaluators are familiar with course content and context; it can be less threatening than supervisor evaluation; and it can have high face validity (Pennington & Young, 1989, pp. 637-638). On the negative side, this method may have low reliability; may involve a conflict of interest resulting in biased reviews, especially if results are used for summative purposes; and criteria are sometimes open to various interpretations. Teachers may also fear damaging their working relationships by evaluating one another (Pennington & Young, 1989, p. 638). (For further discussion of peer review see Christen and Murphy, 1987)

Self-Evaluation

Self-evaluation can be accomplished through such methods as self-reports, self-study materials, self-rating forms, comparison of oneself to one's peers, and videotaping and analyzing one's own teaching. These evaluations can be of value for formative purposes, as they allow for self-reflection and improvement (McGreal, 1983), and encourage professionalism and the establishment of long-term goals for development. Nunan (1988) points out that self-evaluation, when tied to teacher-research, can contribute to curriculum development (p. 147). This method has the added advantage of being inexpensive. However, it suffers from low reliability due to its subjective nature (Pennington & Young, 1989); also, external support and feedback are missing (Barnett, 1983).

Student Evaluations

Student evaluation of teachers is one of the most common methods used in teacher evaluation. Formal student evaluations can be used both summatively (usually as just one component of an evaluation process) and formatively (if done during, rather than at the end of, the course); informal evaluations are used almost exclusively for formative purposes. Aleamoni (1987b) has shown that this form of evaluation can have high reliability and concurrent validity

when developed by an expert, but many forms used do not meet these criteria. Even when forms have been validated, instructors may question their value and may therefore fail to consider them as useful sources of information.

In an ESL or other second language context, there may be special obstacles to overcome. For example, several ESL experts warn of problems that differences in cultural background can cause, such as students being suspicious of the purposes, or feeling pressured by their culture to mark "outstanding" for each item (Saltzer, 1982, p. 96). There may also be comprehension problems due to language limitations, and students may be unfamiliar with scales used in American educational contexts (Pennington & Young, 1989, p. 629). A recent study of ESL student evaluations of teachers in two large ESL programs confirmed these suspicions: there was systematic bias due to ethnic background, level, course content, and attitude toward the course. Thus, using such evaluations in personnel decisions may be unfair (Wennerstrom & Heiser, 1992).

To make student evaluations more reliable and valid for use in ESL contexts, it may be necessary to construct instruments so that factors within the teacher's control are in a separate section from those beyond his or her control; ethnic mix in classes may need to be adjusted for; and teachers may need to be evaluated in a variety of types and levels of courses (Wennerstrom & Heiser, 1992, pp. 283-285). However, a better way may be to use student evaluations of ESL teachers for formative purposes only, emphasizing the use of qualitative feedback obtained from both formal and informal measures. (For further discussion of student evaluation of teachers see Cooper and Miller, 1991; Gaski, 1987; Newstead and Arnold, 1989.)

CURRENT INSTRUMENTS

A good system of teacher evaluation makes use of several different methods and instruments to fulfill the purposes of the evaluation. Moreover, the process ought to have a professional orientation, or an orientation that is geared toward long-term career development, that assesses both teaching and nonteaching aspects of the job, and that includes input from supervisors, students, peers, and individual teachers themselves (Pennington & Young, 1991). In the recent literature, there are several instruments that programs could make use of by adopting them, by adapting them for their particular needs, or by using them as a basis for updating their current instruments.

Instruments for Administrative Performance Reviews

Teacher evaluation should be based on a written job description given to teachers from the outset of their employment. Pennington and Young's (1991) Sample Faculty Standards of Performance makes explicit the program's criteria for effective teaching. It states that these standards clarify teacher responsibilities, make the program's goals explicit, and relate directly to teaching success. It also states that all teachers will be evaluated according to these standards. Under each of the 14 listed standards (e.g., the teacher-student relationship, materials, class activities professional image) is at least one paragraph explaining in detail what is expected (pp. 217-222). Its explicitness and detail, as well as the comprehensiveness and orientation, make it a good model for other programs to look at and possibly adapt.

Pennington and Young (1991) also provide a Format for Annual Teacher Performance Review, which lists the procedures involved in the performance review process (pp. 226-227). Included in the sources of input for the review (Part A) is a list of what subjects and levels were taught; an annual review of activities and accomplishments; biannual couse plans; observation reports; and student evaluation summaries. The areas of evaluation (Part B) are classes; administrative duties; relationships within the program; and professional growth. Under each of these areas is a specification of what they include (e.g., listed under classes are following the curriculum; using materials and equipment; classroom dynamics and atmosphere; and adaptability). The summary of the performance review (Part C) includes strengths as well as weaknesses, and the last step in the review (Part D) is action steps. Teachers should be provided with a copy of this

type of form during orientation, since it makes evaluation procedures clear, and demonstrates to them that input from many sources is considered to be valuable.

White, Martin, Stimson, and Hodge (1991) stress the importance of preparation for performance reviews and provide a form that teachers fill out prior to the annual review (pp. 75-77). For example, teachers should consider their understanding of the purpose of their job and what is expected of them; how well they have done their job; what their strengths and weaknesses are; and what their personal goals are for the next year. Such preparation seems to be a valuable step in the process as is encourages reflection and substantive discussion during the interview portion of the performance review. We feel that this step should be included in any program's evaluation process.

An instrument for teacher observation provided by Pennington and Young (1991, pp. 212-213) lists areas to be included in the observation, and leaves space for comments on each of these. Main headings are preparation; execution; activities; interaction and social climate; teacher's characteristics; and overall assessment of the lesson. Under each of these are specific items to be commented upon. For example, under teacher's characteristics are patience and self-control; confidence; adaptability; voice; use of language; and movement. Some evaluators may appreciate that open-ended comments are called for on this form, and that it covers the most important aspects of teaching, some of which might otherwise be overlooked by observers.

Brown (1995) provides an observation form for administrative classroom visits. Only the five main category names are listed (pace of lesson, teacher presentation, class management, teaching aids, and student production), with two columns next to each category, one for positive aspects and one for suggestions/ideas (p. 196). Although this form by itself would not be suitable for less experienced observers, Brown also provides a list of more specific descriptors that may be used. For example, under student characteristics, Brown suggests that the observer might comment on whether students seem highly motivated and actively involved, and whether error correction is effective (p.197).

Instruments for Peer Review

Brown's (1994) form for peer observation allows for both quantitative and qualitative feedback (pp. 432-434). Specific items listed under each category are rated N/A, 4 (excellent), 3, 2, or 1, and space is left for each of these to be commented on. Major categories are preparation; presentation; execution/methods; personal characteristics; and teacher/student interaction. In the category execution/methods, an example of an item to be rated is the degree to which the teacher was able to adapt to unanticipated situations. While some teachers might not like being "graded," others might appreciate this combination of ratings and comments, especially if it is used as the basis for a post-observation debriefing where reasons for certain ratings and comments can be further explored. If faculty peer observations are used as part of the summative performance review process, then this form may help with, but will certainly not assure, peers' objectivity. It would probably have high face validity with at least some teachers and some administrators.

Instruments for Self-evaluation

A useful self-evaluation instrument is Christison and Bassano's (1984) self-observation form, which is also reprinted in Brown (1994, pp. 435-436). Items are listed by category, and can be rated according to a scale provided. Learning environment includes relationship to students; the classroom; presentation; and culture and adjustment. The category on Individuals includes physical health; self-concepts; aptitude and perception; reinforecement; and development. The category on Activity includes interaction and language. This form is especially appropriate for ESL teachers, since several items deal with concerns specific to teachers of this population. For example, under culture and adjustment, the teachers rate themselves on awaremess that cultural differences affect the learning situation.

Pennington and Young (1991) also provide a Self-Evaluation of Lesson form with questions to ask oneself and space for answering each one (e.g., "Did I give appropriate

feedback?") (pp. 214-215). Teachers more interested in writing comments than in rating specific items would prefer this type of instrument to that of Christison and Bassano, although both instruments could be used in combination.

Instruments for Student Evaluations of Teachers

Pennington and Young's (1991, pp. 208-211) instrument is a 7-point rating scale (6 through 0). The instructions tell students that this questionnaire is meant to give instructors feedback while the course is in progress; therefore, this type of instrument is clearly meant for formative purposes, and can be very useful in helping teachers make appropriate adjustments to their course or to particular aspects of their teaching before students rate them for summative purposes. Brown (1995) offers a sample instrument for student evaluation of ESL teachers (pp. 202-203). Students give a rating value to each item, but also write comments in a separate section. Thus, this form balances the need for an instrument that is easy to format and score, with the need to allow students more freedom in expressing their opinions. Student comments are harder to summarize, but are more meaningful, and therefore useful, to teachers (pp. 200-204).

MAJOR CONSIDERATIONS IN TEACHER EVALUATION

According to Haefele (1993), classroom observation by supervisors, the most common method of teacher evaluation "is in trouble. Evaluation criteria lack validity. Evaluators are untrained. Evaluators consistently award lenient ratings to weak and incompetent teachers. These and other practices . . . clearly indicate that the current teacher evaluation model is seriously deficient" (p. 21). It is our view that these problems are not limited to observations done by supervisors, nor are they insurmountable. With enough planning and information about teacher evaluation, these difficulties can be minimized.

The recommendations that follow are intended to address some of the problems referred to by Haefele (1993). They correspond to the five areas of teacher evaluation that we have already discussed in various ways: the purpose of the evaluation, the content of the evaluation, the development of the instrument to be used, the administration and scoring of the evaluation, and providing feedback to the evaluatee. In addition, a suggestion for integrated evaluation measures is proposed.

Purpose

Teacher evaluation, as mentioned, can have two broad purposes, formative or summative. However, Isenberg (1990) seems to prefer that evaluation be limited to formative uses. She reports that the goal of evaluation is actually to improve teaching, and "should not be used as a threat or to hurt the teacher, but rather as a way to improve the quality of the employee's performance and to facilitate professional growth" (p. 16). A stronger focus on formative evaluations may be desirable, but summative decisions must be made as well. Therefore, while it may be worthwhile to increase the use of formative evaluations to improve teaching quality, summative evaluations also play a crucial role and cannot be ignored. However, improvements can be made in the amount and type of feedback given in evaluations. This will be treated later in this section.

Content

In deciding on the content of the evaluation, the definition of teacher effectiveness must be clearly and explicitly established in order to ensure that all important aspects of performance are included. Content validity becomes an issue here, as discussed by Aleamoni (1981, pp. 118-127). The items included, as well as the format in which the items appear, may affect the outcome of the evaluation. Thus, it is extremely important that the instrument be validated. Herrmann (1987) maintains, based on her study, that both quantitative and qualitative data should be gathered in

order to provide sufficient and varied information. Qualitative data can be particularly useful in formative evaluations since it will allow the instructor to receive substantive feedback.

Feldvebel (1980) concludes that "evaluation based upon a combination of process and product criteria enhances the credibility of the process" (p. 419). In an attempt to identify all of the necessary content of a teacher evaluation, Annadale (1974, p. 45) lists seven areas of teacher effectiveness that should be sampled. These factors include the teacher's preparation and organization, student involvement, the clarity of communication, stimulation (of students), teaching style (e.g., does the style accommodate the students' needs?), exams and evaluations (e.g., are they appropriate and well-constructed?), and learning or self-evaluation. Further suggestions of the same type were made by Simpson and Seidman (1962). To this list one might add teacher qualifications such as knowledge of the content area, and ability to cope with classroom situations. Aleamoni (1987a) outlines recommended components that should be included in college faculty evaluations (student evaluation, self-evaluation, measures of student achievement, peer review, department head assessment, and feedback on all assessment results) and discusses how implementation of this system would lead to more effective instruction.

Instrument

In the process of developing the instrument to be used, great concern must be shown for validity and reliability by utilizing the knowledge and assistance of experts. In addition, both evaluators and evaluatees should be involved in the construction of the evaluation instruments since it is they who best know the tasks and responsibilities associated with the instructional content and context. Root and Overly (1990, p. 35) succinctly suggest that we should "involve key stakeholders in the decision-making process." In fact, Gitlin and Smyth (1990) go so far as to criticize evaluation procedures that are developed without teacher input because this not only conveys the false message that teachers are not experts, but also angers teachers, and creates an atmosphere of teacher distrust of the entire process. "Evaluation thus becomes transformed into a kind of game where winning means finding ways to please the supervisor, and then returning to the same old routine when the door closes" (p. 26).

Administration and Scoring

In order to standardize the administration of evaluations, and thereby improve their reliability, Aleamoni (1981, p. 130) suggests that trained testing personnel be responsible for administering evaluations. A lack of training could lead to the unrecognized misuse of the instruments and of the results, and therefore an invalid evaluation. Furthermore, observers may not be trained in how to appropriately convey otherwise properly obtained data to teachers, and may thereby create serious morale, or even legal, problems. As has been repeatedly noted, reliability and validity are needed for the evaluation to be legally defensible. Haefele (1993, p. 26) mentions that "commitment to and involvement in a rigorous training program for evaluators will improve the validity, reliability, and ultimately the acceptability of the evaluation information." It is fortunate, then, that more interest and resources have been focused on such training in the last decade, as noted by Conley and Dixon (1990).

Evaluations should be administered prior to the end of the course so that improvements in instruction can be made during the course. Aleamoni (1981), for example, cites research indicating that mid-term student evaluations lead to both instructional and student improvement. To be useful, however, such feedback should be given by supervisors in personal consultations (Aleamoni, 1987b, p. 115). As for supervisor evaluation of teachers, the shift in the role of the evaluator from that of observer to that of instructional leader means that post-observation consultations will be much more meaningful to teachers than in the past (Herrmann, 1987, p. 30). No matter what form the evaluation takes, it must be remembered that evaluators need to inform teachers about the purposes, philosophies and procedures involved in order for the process to be valid (Annadale, 1974, p. 80).

As of 1990 for many in the teaching field, evaluator credibility was questionable, and was a possible cause of disillusionment, according to Root and Overly (1990, pp. 37-38). Thus, ensuring proper administration and valid and reliable scoring in the evaluation process are areas that many educational systems need to address more carefully by training those who administer and score the evaluations, both initially and in ongoing workshops or courses. This training should include peer evaluators, especially if their input is used for summative purposes (Pennington & Young, 1989, pp. 637 & 639).

Feedback

The manner in which feedback from the evaluation is given to the evaluatee is important, according to Aleamoni's (1981) research. A method for teachers to tabulate and summarize their own results should be available to teachers so that the tendency of many evaluatees to focus on either the positive or negative extremes is lessened. The main point is that feedback should be easy to understand and interpret. In conferencing formats, teachers should be allowed to explain their perspective on the lesson, and should have the opportunity to discuss any comments or feedback given to them. The content of the feedback is most helpful if the teachers are told what aspects of their teaching seem to be effective, and if they are given information on how to improve their instructional performance (Root and Overly, 1990, p. 34).

Integrated Measures

Since each type of evaluation has its advantages as well as its associated problems, many recent publications suggest that a combination of formats will best serve the needs of both evaluators and evaluatees. In this way, several perspectives can be included in the evaluation, giving balance to the results. This may be particularly important where some measures are suspected of containing error due to bias. The Committee of the Evaluation and Improvement of Teaching at the University of Washington (1982) recommends that student evaluations not be used as the only basis for personnel decisions; rather, student evaluations should be used in conjunction with other instruments such as administrator evaluation, peer review, and self-assessment. Furthermore, they advise that student evaluations be collected from several courses, over an extended amount of time (Wennerstrom & Heiser, 1992, p. 271). The inclusion of both qualitative and quantitative assessment measures is important, as this provides useful information both for evaluating teachers and for helping teachers improve the quality of their teaching (Herrmann, 1987). These varied types of information can be gathered using combinations of formats such as student evaluations used with administrator observations of classrooms, and objective records of professional growth.

One positive way to incorporate many of the suggestions mentioned is to utilize a portfolio format. Such a format allows cooperative decision-making on the content to be included. It also provides an ongoing record of professional development, and a means of maintaining a continuous basis for collecting information. Feedback can be provided in a number of ways, and is not restricted to specifically-scheduled appointments, as feedback conferences tend to be. Portfolios can integrate many perspectives by including student work and evaluations, peer reviews, administrators' comments and official reports, and various forms of self-evaluation. This is particularly interesting in that records such as videotapes can later be reviewed as a whole or in segments. (For further discussion of teacher portfolios see Perkins & Gelfer, 1993; Urbach, 1992.)

In addition, portfolio evaluation may well encourage teachers to participate in activities designed to foster professional growth, such as reflective teaching, classroom-based teacher research projects, and continuing education pursuits. Besides being beneficial to the teachers themselves, these activities can have a direct positive effect on their students, their peers, and the community.

CONCLUSION

In designing a faculty evaluation system, a central concern is determining on what basis teachers will be judged and who will do the judging. We have argued that the criteria decided upon for evaluation should correspond to the job description and standards given to teachers before they even begin teaching. We have also presented a discussion of the criteria, the methods, instruments, and procedures that comprise the evaluation process.

Although each administration must determine which evaluation system best serves their purposes, some elements should be common to all teacher evaluations regardless of context. Most important of these elements is the involvement of teachers in developing all aspects of the evaluation system so that it will be relevant, effective, and beneficial to individual teachers as well as to the program as a whole. Multiple sources of information should reflect the concept that teaching is complex, and the information gathered should provide evidence relevant to the established evaluation criteria. Finally, faculty and administrators should work together to review and improve their program's evaluation processes on an ongoing basis, taking special care to assure a focus that encourages the professional development of faculty members.

THE AUTHORS

Catherine Barrette is a Ph.D. Candidate in the Interdisciplinary Ph.D. Program in Second Language Acquisition and Teaching at the University of Arizona. Her major is Second Language Pedagogy and Program Administration and her minor is Second Language Analysis. She can be reached at cdonnell@ccit.arizona.edu

Elaine Morton is a Ph.D. Candidate in the Interdisciplinary Ph.D. Program in Second Language Acquisition and Teaching at the University of Arizona. Her major is Second Language Pedagogy and Program Administration and her minor is Second Language Processes and Learning. She can be reached at emorton@ccit.arizona.edu

Anjel Tozcu is a Ph.D. Candidate in the Interdisciplinary Ph.D. Program in Second Language Acquisition and Teaching at the University of Arizona. Her major is Second Language Pedagogy and Program Administration and her minor is Second Language Use. She can be reached at aintab@wolf.ncat.edu

REFERENCES

Akpe, C. S., & Igwe, G. 0. (1992). Towards an effective and reliable evaluation of college teaching practice. *Studies in Educational Evaluation, 18*, 221-226.

Aleamoni, L. M. (1978). Development and factorial validation of the Arizona Course/Instructor Evaluation Questionnaire. *Educational and Psychological Measurement, 38*, 1063-1067.

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 110-145). Beverly Hills: Sage.

Aleamoni, L. M. (1987a). Evaluating instructional effectiveness can be a rewarding experience. *Plant Disease,71*, 377-379.

Aleamoni, L. M. (1987b). Student rating myths versus research facts. *Journal of Personnel Evaluation in Education, 1*, 111-119.

Annadale, A. D. (1974). *The development and validation of an instrument to measure teaching effectiveness*. Unpublished doctoral dissertation. The University of Arizona, Tucson, AZ.

Barnett, M. A. (1983). Peer observation and analysis: Improving teaching and training TAs. *ADFL Bulletin, 15* (1), 30-36.

Benton, S. E. (1982). *Rating college teaching: Criterion validity studies of student evaluation-of-instruction instruments*. Washington, D.C.: American Association for Higher Education.

Brown, H. D. (1994). *Teaching by principles: An interactive approach to language pedagogy*. Englewood Cliffs, NJ: Prentice Hall Regents.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development.* Boston, MA: Heinle and Heinle.

Carroll, J. G. (1981). Faculty self-evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 180-200). Beverly Hills: Sage.

Christen, W. L., & Murphy, T. J. (1987). Inservice training and peer evaluation: An integrated program for faculty development. *NASSP Bulletin, 71* (500), 10-18.

Christison. M., & Bassano, S. (1984). Teacher self-observation. *TESOL Newsletter* (August), 17-19.

Conley, D. T., & Dixon, K. M. (1990). The evaluation report: A tool for teacher growth. NASSP *Bulletin, 74* (527), 7-14.

Cooper, S. E., & Miller, J. A. (1991). MBTI learning style-teaching style discongruencies. *Educational and Psychological Measurement, 51* (3), 699-706.

Delamere, T. (1986). On the supervision and evaluation of instruction. *System, 14*, 327-333.

Fanselow, J. F. (1988). "Let's see:" Contrasting conversations about teaching. *TESOL Quarterly, 22*, 113-130.

Feldvebel, A. M. (1980). Teacher evaluation: Ingredients of a credible model. *The Clearing House, 53*, 415-420.

Flanders, N. A., (1970). *Analyzing teaching behavior.* Reading, MA: Addison-Wesley.

Gaski, J. F. (1987). On "construct validity of measures of college teaching effectiveness." *Journal of Educational Psychology, 79* (3), 326-330.

Gitlin, A., & Smyth, J. (1990). Educative possibilities in teacher evaluation: Two alternatives. *NASSP Bulletin, 74* (527), 25-32.

Haefele, D. L. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education, 7*, 21-31.

Herrmann, B. A. (1987). Effective teacher evaluation: A quantitative & qualitative process. *NASSP Bulletin, 71* (503), 23-30.

Isenberg, A. P. (1990). Evaluating teachers--Some questions and some considerations. *NASSP Bulletin, 74* (527), 16-18.

MacGinitie, W. H. (1993). Some limits of assessment. *Journal of Reading, 6*, 556-560.

Magnusen, K. 0. (1987). Faculty evaluation, performance, and pay. *Journal of Higher Education, 58* (5), 516-529.

Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness. *Journal of Higher Education, 64* (1), 1-18.

McGreal, T. L. (1983). *Successful teacher evaluation.* Alexandria, VA: Association for Supervision and Curriculum Development.

Nerenz, A. G., & Knop, C. K. (1982). A time-based approach to the study of teacher effectiveness. *Modern Language Journal, 66*, 243-254.

Newstead, S. E., & Arnold, J. (1989). The effect of response format on ratings of teaching. *Educational and Psychological Measurement, 49* (1), 33-43.

Nunan, D. (1988). *The learner-centered curriculum: A study in second language teaching.* Cambridge: Cambridge University Press.

Nunan, D. (1989). *Understanding language classroams: A guide for teacher- initiated action.* New York: Prentice Hall.

Pennington, M. C., & Young, A. L. (1989). Approaches to faculty evaluation for ESL. *TESOL Quarterly, 23*, 619-646.

Pennington, M. C., & Young, A. L. (1991). Procedures and instruments for faculty evaluation in ESL. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 191-205). Washington, D.C.: NAFSA: Association of International Educators.

Perkins, P. G., & Gelfer, J. I. (1993). Portfolio assessment of teachers. *The Clearing House, 66* (4), 235-237.

Pope, C. A. (1990). Indirect teaching and assessment: Are they mutually exclusive? *NASSP Bulletin, 74* (527), 1-5.

Root, D., & Overly, D. (1990). Successful teacher evaluation: Key elements. *NASSP Bulletin, 74* (527), 34-38.

Saltzer, M. G. (1982). The evaluation of an intensive English program. In R. P. Barrett (Ed.), *The administration of intensive English language programs* (pp. 89-97). Washington, D.C.: National Association for Foreign Student Affairs.

Sheal, P. (1989). Classroom observation: Training the observers. *ELT Journal, 43*, 92-104.

Simpson, R. H., & Seidman, J. M. (1962). *Student evaluation of teaching and learning: Illustrative items for teacher self-evaluative instruments.* Washington, D.C.: The American Association of Colleges for Teacher Education.

Tracey, W. R. (1978). Teacher evaluation: Another perspective. *The Clearinghouse, 51*, 240-242.

Travers, R. M. W. (1981). Criteria of good teaching. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 14-22). Beverly Hills: Sage.

Troyer, M. E., & Pace, C. R. (1944). *Evaluation in Teacher Education.* Washington, D.C.: American Council on Education.

Urbach, F. (1992). Developing a teaching portfolio. *College Teaching, 40* (2), 71-74.

Wennerstrom, A. K., & Heiser, P. (1992). ESL student bias in instructional evaluation. *TESOL Quarterly, 26*, 271-288.

White, R., Martin, M., Stimson, M., Hodge, R. (1991). *Management in English language teaching.* Cambridge: Cambridge University Press.

Wood, D. R. (1992). Teaching narratives: A source for faculty development and evaluation. *Harvard Educational Review, 62* (4), 535-550.