CHAPTER 16

EVALUATION OF TEACHING

It is natural to want to know how well one has done on a given task. In its simplest form, evaluation of teaching allows an instructor to obtain this feedback. Once collected, the data can be used to help the instructor improve the course, compare instructors, reward or punish the instructor, or inform potential students. Since improvement of teaching without this feedback is unlikely, we are in favor of this use of teaching evaluations. Unfortunately, the evaluation of teaching has become embroiled in controversy, partially because of the other uses of evaluations.

In this chapter we will start with a discussion of formative and summative evaluations and the objectives of each; then we will consider the validity of student evaluations, correlations with other methods, and extraneous variables which affect student evaluations. Since student evaluations are only one of many procedures which have been used for evaluating teaching, we next discuss the various other methods.

Many professors in psychology and education have devoted their careers to studying the evaluation of teaching. Although many questions remain, there is a large body of scientifically valid knowledge about the subject. We intend to tap into this knowledge so that the reader can intelligently discuss the issues surrounding the evaluation of teaching. This background information will give the reader a distinct advantage over most engineering professors who discuss these issues on the basis of ancedotal evidence and biases.

16.1. FORMATIVE AND SUMMATIVE EVALUATIONS

Essentially, a course can be evaluated at any time during or after the semester or term. Evaluations made during the course, called formative evaluations, are aimed at eliciting

comments from students so that the professor can make in-course corrections. These evaluations can be as simple as passing out comment cards and asking the students to respond anonymously to two questions such as:

What do you like about this course?
What about this course could be changed to improve your learning?

They are useful if the professor changes things that are not working. If, for example, the comments reveal that the TA is not available during office hours, the professor can take steps to correct this problem early in the semester. The evaluations can also allow the professor to do something he or she wants to do, but which might not go over well without the empowerment of student comments. For example, if one or two students are monopolizing the professor's time during questions and discussion, other students will likely complain on the comment cards. The professor can then say in a positive sense that he or she has been asked to involve more students in the discussion or questions.

There are other types of formative evaluation. Chatting with students informally during the semester often points out what is or is not working. Formal weekly meetings with a group of students representing the class is another way of obtaining useful feedback during the semester. Chatting with the TA can also be illuminating since TAs often have a good idea of what is or is not working. Critically evaluating the results of quizzes or tests may show that certain critical concepts have not been learned. The professor may want to adjust the syllabus to provide more time for these concepts. Watching the students' nonverbal behavior and asking them if they understand is also a type of formative evaluation which can be used in every class period.

Summative evaluations, which are done at the end of the course or well after the course is over, are used for a variety of purposes, some of which are controversial (see Sections 16.2 and 16.4). Of course, summative evaluations provide feedback to the professor. Since professorial self-evaluations are often very high (Centra, 1980), student evaluations can provide a salutary dose of reality. When the professor has done a good job, the feedback is a welcome pat on the back. Summative follow-up evaluations by alumni can also provide feedback as to what course material has proven to be particularly useful in industry (see Section 16.4).

Summative student evaluations can also be helpful in instructor and course improvement. The more specific the comments, the more useful they are for course improvement. Answers to very general rating questions such as "This is one of the best courses I have ever taken" are not useful for course improvement. Questions on the textbook, handouts, availability of help, homework, tests, lectures, and so forth, can provide the professor with specific areas needing improvement. Based on dissonance theory (when the person's self-evaluation and the feedback received from others differ, dissonance is generated and the person reacts to reduce this dissonance), professors should act to improve their teaching based on student ratings. Unfortunately, many studies have shown little or modest improvement in teaching resulting from the use of course evaluations *alone* (Aubrecht, 1979, 1981; Centra, 1980; Lowman, 1985). A meta-analysis by Cohen (1980) shows that there is improvement, but it is modest.

What does work to improve teaching? For a start, specific questions on student ratings coupled with consultation with another professor (Aubrecht, 1981; Eble, 1988; McKeachie, 1986, 1990). Without a consultant most professors either rationalize the ratings or "just try harder." The consultant helps the professor focus on an action plan to solve the problems pointed out in the ratings. This person can make specific suggestions of what to try and can also be supportive. A specific development plan with informal follow-ups can be developed for the remainder of the semester or for the next semester. Since professors are busy and have many obligations in addition to teaching a specific course, we recommend that the consultant be an interested professor in his or her own department. Then the consultant will understand the constraints the professor is acting under and will not make recommendations which are impossible. McKeachie (1986) suggests that there is no reason to wait until the end of the semester to administer the evaluation form. The student evaluation can be useful for course improvement in the current semester if it is administered from the third to the fifth week of the semester.

Student evaluations, whether formative or summative, are useful because they improve student morale. The chance to register an opinion is helpful even if no one pays any attention. Of course, if it is clear that someone is paying attention and the instructor responds to the comments and improves the course, then student morale will improve even further (Abbott et al., 1990). Although it would be manipulative to give students an opportunity to evaluate courses merely to increase student morale, the increase in student morale when evaluations are used for other purposes is obviously a side benefit.

Administrative use of student evaluations tends to be quite controversial (Eble, 1988; Johnson, 1988; Lowman, 1985), especially when salary, promotion, and tenure decisions are involved. The first problem is that student evaluations are often not well administered. It is not unheard of for professors to hand out the evaluations and then to throw away poor evaluations before turning them in for scoring. A uniform administration procedure must be used to avoid this or other abuses (see Section 16.2.2). One possible solution is to use a separate rating form for administrative purposes and to administer it in a senior seminar course (Milligan, 1982). Second, many professors do not trust the reliability or validity of student evaluations. This issue can be partly put to rest with scientific data (see Section 16.3). Unfortunately, if the administrator using the evaluations does not understand the effect of extraneous variables, the evaluations can be misused. For example, evaluations of professors in classes with less than fifteen students tend to be quite high. This needs to be taken into account when professors are compared. A related problem is that the specific questions which are so useful for course improvement are not useful for overall administrative evaluations (Centra, 1980). Only the overall course and instructor ratings are useful for this purpose since the overall ratings have the highest correlations with student learning. To avoid inadvertent abuses, only the overall ratings should be sent to administrators and promotion committees. The alternative of a separate rating form for administrative use only would also solve this problem. A fourth problem is that few professors are uniformly excellent or uniformly poor in all types of courses (Murray et al., 1990). Poor ratings may only represent poor casting of the professor in a course. What types of courses a professor can teach well is obviously useful information, but using student ratings in a single course is not a fair procedure for setting raises or deciding on promotions. Ratings over a long time period for a large number of courses are needed.

Evaluation of teaching for administrative use by faculty or chair visits to the classroom are even more controversial than the use of student ratings. Since ratings based on visits by professors not trained in the evaluation of teaching tend to be much less reliable than student ratings, this practice should not be used for administrative purposes. (Faculty visits can be useful for course improvement; see Section 16.4.)

A final use of student ratings is as information for other students who are potential consumers of the courses (Canelos and Elliott, 1985; Marsh, 1984). Some universities have a long tradition of student-run evaluations which are then published in student guides. There is no doubt that these guides do have an effect on the elective courses which students sign up for. The aim of informing the consumer of what an instructor and course will be like is probably laudable. Unfortunately, student-run ratings and guides may be poorly controlled (and in effect uncontrollable). It is not unusual for some of the guides to be extremely biased, particularly during periods of political upheaval. Engineering courses are usually not heavily represented in these guides since few engineering students join these student groups and since few engineering courses are electives.

16.2. METHODS FOR DOING STUDENT EVALUATIONS

Since student evaluations are now the most common method for evaluating instruction, we will focus on them in this section and in Section 16.3. However, student evaluations by themselves cannot completely evaluate instruction; thus, they should be used in conjunction with other evaluation methods (see Section 16.4).

16.2.1. Types of Student Evaluations

If the purpose of the course evaluation is entirely feedback to the instructor for the purpose of course improvement, then informal evaluation procedures can be used. Both formative and summative evaluations can be made with comment cards, either with or without cues to the students on what to focus on. If there are specific questions of interest, the professor can generate a student questionnaire (Cook, 1975). But for administrative use or for research purposes, professor-generated questionnaires and comment cards are not suitable.

For administrative purposes, global questions on teaching effectiveness should be used since they have the highest correlations with student achievement (Centra, 1980). A simple alternative is to have all the seniors rate the professors on a scale from 1 to 5. Milligan (1982) suggests doing this for each professor regardless of the number of different courses the student has taken from that professor. Since most professors cannot teach all courses with equal skill (see Section 16.3.3), it is probably better to do this evaluation course by course for each teacher. There is an advantage to separating the course improvement and administrative functions of student evaluations, since professors are more likely to use formalized course evaluations if they know they will not be used by the administration.

TABLE 16-1 COMMERCIALLY AVAILABLE FORMS FOR STUDENT EVALUATION OF FACULTY (JOHNSON, 1988)

Form	Comments	Source
Cafeteria	Allows instructor to choose 40 from 200 items. Five (5) overall core items are automatically added. Room for comments. Five (5) point scale.	Center for Instructional Services Purdue University STEW B14 West Lafayette, Indiana 47907
Course Instructor Evaluation Questionnaire (CIEQ)	Up to 63 items instructor generated with 7 on student backgrounds. Has open-ended items. Four (4) point scale.	Division Education Foundation and Admin., College of Education University of Arizona P.O. Box 302 Tucson, Arizona 85721
Instructional Assessment Form (IAS)	General evaluation, diagnostic feedback, information about students and course. Instructor can add items. Can include open-ended items. Six (6) point scale.	Educational Assessment Ctr. University of Washington 453 Schmitz Hall P.O. Box 30 1400 N.E. Campus Parkway Seattle, Washington 98195
Instructional Development and Effectiveness Assessment System (IDEA)	Variable number items. Can be instructor-generated. Instructor and course characteristics, evaluate progress towards course objectives, self-rating by students. Can use open-ended questions. Five (5) point scale.	Center for Fac. Evaluation and Development Kansas State University 1627 Anderson Avenue Box 3000 Manhattan, Kansas 66502
Student Instructional Report (SIR	Thirty-nine (39) items, can use instructor-generated, no open-ended. Five (5) student background items. Instructor and course characteristics, course and instructor variables. Four (4) point scale.	Educational Testing Serv. (ETS) ETS College and University Prog. Princeton, New Jersey 08541-0001

Many universities use formalized course evaluation procedures which are often administered by either a separate learning center or a student organization. The forms used are usually machine-scorable, multiple-choice questionnaires with space available for student comments. The students usually rank a variety of questions on 4- to 7-point scales. Usually both specific items such as "The textbook was well written and understandable" and global ranking items such as "Overall, this course ranks highly" are included in the questionnaire. The forms may allow for instructor selection of items from a large pool, and it may be possible for the instructor to add additional items. Marsh (1984) notes that since good instruction can have many dimensions, the forms must be multidimensional; that is, many different aspects of instructional ability need to be considered.

A large number of course evaluation forms have been developed and are available for a nominal fee. Some of them are listed in Table 16-1. Johnson (1988) gives two samples of the

available forms. Jakubowski (1982) shows a form generated following the comments of a student panel. Janners and Tampas (1986) discuss the development of a form locally so that the faculty will accept its use. They present their final result. Fowler (1978) shows both a multiple-choice form and a form with open-ended questions. A detailed questionnaire for formative evaluations has been developed by Davis and Alexander (1976). Marsh (1984) notes that if student evaluations are to be used for research purposes, the form needs to be carefully designed. Many of the commercially available forms are adequate, and no form has been shown to be superior, which is why many different forms are in use.

16.2.2. Administration of Student Evaluations

Several studies have shown that the way student evaluation forms are administered can affect student ratings (Aubrecht, 1979; Centra, 1980; Marsh, 1984). Professors who make verbal comments requesting high rankings because of their importance in promotion and tenure decisions may well get higher rankings, particularly if the comments are subtle instead of blatant. There is also a built-in bias if the professor is present when the students fill out the evaluation forms. In addition, professors, like students, are subject to human frailty, and both have been known to cheat occasionally.

To avoid these problems a uniform procedure for administering student evaluations should be used throughout the department. The professor should not be present when students are filling out the forms. A trustworthy administrative assistant, TA, or even the department chair should administer the evaluations. A standard procedure such as the following should be followed by this person:

1 Bring in the forms and pencils needed.

2 Announce to the class why he or she is there and state that it is departmental policy that the professor not be present.

3 Describe the purpose of the forms, state what they will be used for, and note that evaluations are important and need to be done carefully.

4 Pass out the forms and pencils.

5 Give simple instructions. Be sure to note that 1 is high (or low).

6 When all the students are finished, collect the forms and put them into an envelope. Seal the envelope.

7 Deliver the envelope to the agency which scores the forms.

What should be done with the results of the evaluations once they have been scored is somewhat controversial. Certainly they should be provided to the professor for course improvement. Professors should be encouraged but not forced to discuss their evaluations with another professor or an instructional consultant. They should also be encouraged to discuss the ratings and an improvement strategy with the class since this increases the students' satisfaction (Abbott et al., 1990). The use of voluntary evaluations for administrative purposes

can cause problems if norms are reported. Since those who volunteer are mainly professors who are most interested in teaching and who are good at teaching, the norms are skewed to high rankings. For administrative uses a required rating of all the professors in the department is preferable.

16.3. STUDENT EVALUATIONS: RELIABILITY, VALIDITY, AND EXTRANEOUS VARIABLES

Many faculty members complain that student evaluations do not mean anything, arguing that they are not reliable, that students can be bought with grades, that the ratings are not valid, that alumni, not students, should do the rating, and so forth. Unfortunately, engineering professors who would never dream of doing an engineering design without data are willing to complain about student evaluations with no data. In this section a sampling of the available scientific data which allows one to discuss these complaints rationally will be presented. Before discussing the questions of reliability, validity, and extraneous variables in detail, we will note that the complaints are often somewhat misplaced. Students are generous evaluators. For example, in a study only 11 percent of 852 engineering classes were rated as below average (Centra, 1980).

16.3.1. Reliability of Student Ratings

Reliability of student ratings means that they are consistent for whatever it is they are measuring. The internal consistency of student ratings is quite good and becomes excellent as the number of students doing the rating increases. For the IDEA rating system Aubrecht (1979) reports the following correlation coefficients:

$r = 0.69$ (ten students)
$r = 0.81$ (twenty students)
$r = 0.89$ (forty students)

For the SEEQ rating system Marsh (1984) reports correlation coefficients that are slightly higher:

$r = 0.6$ (five students)
$r = 0.74$ (ten students)
$r = 0.90$ (twenty-five students)
$r = 0.95$ (fifty students)

Thus, the internal consistency (the agreement of students in the same class) is quite high.

A second measure of reliability is stability. Are the raters stable over time and are the professors stable over time? The correlation coefficient for students in 100 classes when they were asked to rate the class after it was over and at least one year later was $r = 0.83$ (Aubrecht, 1981; Marsh, 1984). When the same instructor was evaluated for the same course but in different years (which means different student raters), the correlation coefficients varied from $r = 0.62$ to $r = 0.89$ with a mean value of $r_{mean} = 0.74$ (Marsh, 1984; Murray et al., 1990). Thus, we can conclude that both student raters and professors teaching the same course are stable.

There is no reason to believe that professors will be equally proficient at teaching all courses. When the same instructor was rated in the same year in different courses, the correlation coefficients varied from $r = 0.33$ to $r = 0.55$ with a mean value of $r_{mean} = 0.42$ (Marsh, 1984; Murray et al., 1990). These correlation coefficients are significantly lower than those obtained for the same instructor teaching the same course. This result is discussed in more detail at the end of Section 16.3.3.

16.3.2. Validity of Student Ratings

Validity means that student ratings are measuring what they are supposed to be measuring. Do student ratings actually measure teaching quality? This is a much harder question to answer than questions of reliability, but sufficient research reports are available to give an affirmative answer.

Critics of student ratings often claim that student achievement is the outcome that we should study. Do student ratings correlate with student achievement? There is broad agreement in the literature that a reasonably strong positive correlation exists between student achievement and student ratings (Aubrecht, 1979; Centra, 1980; Cohen, 1981; Greenwood and Ramagli, 1980; McKeachie, 1986, 1990; Marsh, 1984). The conclusive study was the meta-analysis of Cohen (1981) who looked at all available studies relating student achievement and student ratings in courses with multiple sections taught by different instructors. The global ratings were highly correlated with the final examination scores. Cohen (1981) found correlation coefficients of $r = 0.50$ based on questions about instructor skill, $r = 0.47$ based on questions about the global rating of the course, and $r = 0.43$ based on questions about the global rating of the instructor. Thus, sections where students learned more rated the instructor and the course higher than sections where students learned less.

Student ratings also have modest positive ratings when compared to other methods of evaluating instruction. The correlation coefficients between student ratings and ratings by professors ranged from $r = 0.60$ to $r = 0.70$ if the professor had not visited the classroom (Aubrecht, 1979; Marsh, 1984). The correlation between student ratings and administrator ratings where the administrator had not visited the classroom was $r = 0.47$ (Aubrecht, 1979). If the colleague had visited the classroom before rating the instructor, then the correlation coefficient with student ratings was $r = 0.20$ (Aubrecht, 1979; Marsh, 1984). This number is low partially because the reliability of ratings based on colleague visits is low (see Section 16.4). When professors did not visit a colleague's classroom, they apparently based at least

part of their ratings on discussions with students. Thus, these ratings correlate significantly higher than those based on visits.

The correlations of professors' self-rating of their teaching ability with student ratings has been extensively studied. Correlation coefficients between student ratings and a general instructor self-rating are about $r = 0.19$ (Greenwood and Ramagli, 1980). When the instructors do a self-rating for a specific course, then the correlations with student ratings are significantly higher, $r = 0.45$ to $r = 0.49$ (Marsh, 1984). Most professors rate themselves higher than the students do, and about 30 percent of the time significantly higher.

Factor analysis has been used to determine what students are rating. The results of these studies show that students do not just give a single global rating but include several factors. Aubrecht (1979) states that a typical breakdown of factors with the most important factor first is:

1 Skill. Interesting presentation, intellectual stimulation, clarity.
2 Rapport. Concern for students, classroom interaction.
3 Structure. Organization, course preparation.
4 Difficulty. Amount of work demanded.

A similar but more detailed list of seven factors is given by Marsh (1984):

1 Learning and value. Challenge, subject interest, amount of material learned.
2 Enthusiasm. Interest, humor.
3 Organization. Objectives, clear explanation.
4 Group interaction.
5 Individual rapport. Provides help and answers questions.
6 Breadth of coverage.
7 Examinations and grading.

Wilson's (1972) list includes the first five factors given by Marsh.

Higgins et al. (1991) prepared a list of the top ten characteristics of instruction by asking their engineering students to generate such a list. Their list in order of importance is:

1 Organized and prepared.
2 Simple, straightforward instruction with complete examples.
3 Good communication and pronunciation.
4 Real-life applications and analogies.
5 Sound knowledge of subject matter.
6 Open to questions during and after class.
7 Goals clearly stated at beginning.
8 Interested in subject.
9 Lots of examples.
10 Logical order and avoids tangents.

The procedure used to create this list is significantly different from factor analysis. In addition, the list was generated from a rather small number of engineering students instead of a large number of students from many areas as was done for the factor analysis. The engineers were somewhat more pragmatic and wanted examples, but other comments were similar to the factor analysis lists.

These analyses are further proof of the validity of student ratings. Students have rated by reasonable criteria for good teaching. These ratings also agree with the two-dimensional model of good teaching which was presented in Chapter 1.

16.3.3. Effects of Extraneous Variables on Student Ratings

Critics attack the validity of student ratings because of the effect of extraneous variables. They state that ratings are affected by the time at which the class is taught, who is taught, the grades given, the class size, the type of course, the age and gender of the professor, and so forth. This attack is partially correct since extraneous variables do affect student ratings, but the effect is usually quite small and is not enough to make a good teacher look poor, or vice versa (McKeachie, 1990). In this section we will explore what Marsh (1984, p. 730) calls "the witch hunt for potential biases in students' evaluations."

Initial Student Motivation and Expectations. Students who expect a course to be good often find this to be true (McKeachie, 1986). The correlation between the student's initial liking for the subject and the student's rating of the course at the end of the semester range from $r = 0.42$ to $r = 0.49$ which are quite high (Aubrecht, 1979, 1981). Student enthusiasm and prior interest account for much of the background or extraneous variable effect (Marsh, 1984). Initial student motivation is such an important variable that in the IDEA system for teacher evaluation, initial student motivation is used in combination with class size to establish norm groups for comparison purposes (Aubrecht, 1979).

Class Size. The second most important extraneous variable is class size, but the correlation coefficients are significantly less than for initial student motivation. When there are fewer than fifteen students in a class, the ratings are significantly higher than they are otherwise. Students enjoy the close personal contact with the professor and with other students, which is almost automatic in classes this small (Centra, 1980). As the class gets larger the ratings decrease and the correlation coefficients obtained are generally from $r = -0.10$ to $r = -0.30$ (the correlation is negative since ratings are smaller for larger classes) (Aubrecht, 1979; Koushki and Kuhn, 1982). For very large classes (more than 200 students), several studies show that ratings go back up (Marsh, 1984; Koushki and Kuhn, 1982). This may occur because departments assign their best teachers to large classes. Note that some studies have found no effect of class size in engineering courses (Canelos and Elliott, 1985).

Academic Field. There are small but significant effects based on the academic discipline when all other variables are controlled. Aubrecht (1979) reported that humanities, fine arts, and language had slightly higher rankings than social or physical sciences, mathematics, and engineering. Koushki and Kuhn (1982) found that at Clarkson University the arts and sciences and industrial distribution had slightly higher ratings than either engineering or management.

Although there is not complete agreement between these studies, they do agree that engineering students give ratings at the low end of the spectrum. Thus, cross-field comparisons are somewhat difficult.

Course Type. Engineering professors commonly believe that laboratory courses receive low ratings. Kuriger (1978) found that this was indeed true and that these courses had much lower ratings than classes dispensing theory. Kuriger (1978) also found that engineering elective courses had better ratings than required courses in the engineering discipline, which had higher rankings than core engineering classes. Seniors and graduate students rated classes slightly higher than other students even when the type of class was the same (Kuriger, 1978). Koushki and Kuhn (1982) also found that electives and courses in the discipline had higher ratings than core courses, but they did not observe a difference between elective and required courses in the discipline. The hours that the class meets also makes a small difference, with classes meeting at the convenient times of midmorning and midafternoon receiving the highest rankings (Koushki and Kuhn, 1982).

Grades and Course Workload. A common criticism is that professors can buy ratings by requiring very little work and by easy grading. The first hypothesis is clearly wrong. Studies show that students rate courses with high workloads higher than courses with low workloads (Marsh, 1984). The effect of grades is much more complex. Grades earned previously from the same instructor do not affect ratings in the course (Canelos and Elliott, 1985). Although Kuriger (1978) found a negligible correlation between grades and ratings in engineering courses, pooled studies over many classes show correlation coefficients between expected grade and ratings ranging from $r = 0.1$ to $r = 0.3$ (Aubrecht, 1979). However, one needs to be careful not to confuse correlation with causation. When the studies are controlled for prior interest in the subject and for the effect of workload in the course, most of the correlation disappears (Aubrecht, 1981; Marsh, 1984). What remains is mainly from students who are receiving A's. Marsh (1984) discusses three possible hypotheses for the remaining slight effect of expected grade on course rating. These hypotheses are:

1 Grading leniency. The students rate the course higher because they expect a grade higher than they have really earned. There is no empirical evidence for this hypothesis.

2 Validity. Better learning in the course as illustrated by a higher grade leads to a better rating of the course.

3 Student characteristics. Students who earn better grades have some characteristic which leads them to rate the course higher. This is correlation without causation.

Professors. A large number of professor effects have been studied. First, Kuriger (1978) found that professors who had won teaching awards received significantly higher rankings than professors who had not. This is no surprise and represents another sign of the reliability of student ratings. Kuriger (1978) also found that professors received better ratings than American TAs who had higher ratings than foreign TAs. Presumably, the professors are more experienced. However, younger professors do better than older professors (Canelos and Elliott, 1985). By a slight amount, associate professors received the highest ratings, but this disappeared when only electives were considered (Kuriger, 1978). Students do react posi-

tively to very expressive teachers, and these teachers may get overly generous ratings (McKeachie, 1990). However, one of the items that students consider a constituent of good teaching is enthusiasm, and expressiveness is interpreted as enthusiasm. Neither the gender of the instructor nor the knowledge of the subject matter affects the ratings (McKeachie, 1990). The question of how the professor's research affects teaching ability and ratings has been extensively studied and has proven to be complex. Research either has no effect or a slight positive effect on ratings. This issue is discussed in detail in Section 17.3.

Instructor Personality. Murray et al. (1990) did a very interesting study on the interactions between the professor's personality and the type of course. The most important conclusion from their study is that few professors are good teachers in all types of courses and few teachers are poor teachers in all types of courses. Casting the professor in the appropriate type of course is important. The authors suggest that professors should determine what type of course they do well in and then stay with that type of course as much as possible. The three general categories of courses which were clearly different were introductory and general courses which were held in large lecture halls, junior- and senior-level electives which were much smaller discussion classes, and methodology courses which were very work-intensive. Professors who were extroverts, yet compulsive enough to handle the details of large classes, received high ratings in the large lecture classes. Professors who were extroverted, friendly, and supportive yet flexible received high ratings in the discussion classes. Ambitious, competent, hardworking, and confident professors did well in the methodology courses. The only personality trait which correlated with high ratings in all categories of courses was leadership, which they defined as taking initiative and getting things done. Note that this study involved psychology courses and may not generalize to other fields. In a separate study Sherman and Blackburn (1975) found that instructor pragmatism was positively related to ratings in natural science courses but not in courses in humanities or social sciences. Instructor amicability was related to ratings in humanities but not in natural or social sciences. From this one could hypothesize that pragmatic instructors would receive higher ratings in engineering courses.

16.3.4. Can a Professor "Buy" Student Ratings?

Yes, a professor can "buy" student ratings by two different methods. First the professor can load all the extraneous variables in her or his favor. Thus, the professor could arrange to teach a small, nonlaboratory, elective class to seniors and graduate students. The course would be scheduled at a convenient time, and the TA would be from the United States. If possible, the students would be initially interested in the material. The professor would give A's to all the students on the A-B border. This set of conditions can buy a slightly higher rating, but it cannot turn a poor teacher into a good one.

The second approach is to present material clearly and communicate with the students. Organize the material and give clear objectives. Follow a logical presentation scheme with a minimum of tangents. Present many examples and real-life applications. Cultivate a pragmatic, let's-get-things-done attitude. Show enthusiasm, interest, and a love for the subject.

Stimulate the students intellectually and have a significant breadth of coverage. Be available for questions both in and out of class. Have a sense of humor. Use a good textbook which is integrated into the course. Arrange matters so that the workload is high, but not unreasonably so. Have fair examinations and a clearly defined grading system. Encourage group interactions both within and outside the class. Develop a team concept with the students—a team whose job it is to learn the material. Keep the students active and incorporate a variety of modes of presentation. If all these things are done, then the professor will have done a good job and will have earned the high ratings he or she will receive.

16.4. OTHER EVALUATION PROCEDURES

Student evaluations, though useful, are neither the sole nor the best way to evaluate a course. They miss, for example, the richness of ideas which can be obtained with interview techniques, and students are also often not qualified to evaluate content. A combination of techniques can make up for the deficiencies of student ratings.

Student interviews can be a much richer source of information than student ratings, but they are time-consuming. Interviews should not be done by any of the professors who are being evaluated. If the department chair can arrange to interview all the graduating seniors, a significant amount of information can be obtained about the performance of professors, the curriculum, and miscellaneous items. Except in regard to courses the students have taken recently, the information is not likely to be specific enough to help professors improve courses. Thus, the interviews should supplement course evaluations. For valid information to be obtained on professors, a high percentage of the students need to be interviewed; otherwise, only students with complaints may come in. Although the main advantage of interviews is that students have the freedom to bring up whatever they want, some structure helps control the time and ensures important topics are covered. Setting a time limit in advance is useful since it helps the interviewer structure the interview and control time.

An alternative to individual student interviews which takes much less time is the Small Group Instructional Diagnosis (SGID) method (Abbott et al., 1990) in which a facilitator and the instructor first meet to discuss the course. The facilitator then meets with the class in the absence of the instructor and forms small groups which discuss the strengths of the class, areas requiring change, and recommendations for change. Each group reports to the class, and the facilitator collects and summarizes the reports for the class. He or she then clarifies the ideas until the class agrees that the summary is accurate. This class meeting can take place in a single fifty-minute period. The facilitator and the instructor then meet to discuss the students' concerns and recommendations. A strategy for improving teaching is developed. The instructor returns to class and extensively discusses the facilitator's report and the proposed improvement strategy. Of the methods tried, students preferred the group interview procedure to the use of standardized rating forms. With either the group interview or standardized rating forms, students were more satisfied when the instructor responded extensively to the student evaluations (Abbott, et al., 1990).

Self-ratings by instructors are useful for course improvement, although the correlations with student ratings are low. Since many faculty rate themselves high, with 30 percent significantly higher than the students' evaluations, self-ratings should be used as only one part of the course evaluation system. Instructors are more realistic in their self-ratings when they focus on a specific course. Use of some type of questionnaire such as the course evaluation guide developed by Lindenlaub and Oreovicz (1982) helps to ensure that the instructor has not missed any important areas. Course improvement is highest when the self-evaluation is discussed with a supportive but critical consultant. This is particularly true regarding the pace of the course and the workload. Natural science professors typically underestimate the pace and workload (Greenwood and Ramagli, 1980), and engineering professors probably do also.

Consultation can be used with any of the other techniques such as student ratings. Course improvement is much more likely if the ratings are shared with a consultant, probably because it is much harder to avoid the signals that some improvement is needed. An unstructured conversation—letting the person just talk about teaching—can be very useful in providing insights (Elbow, 1986). The unstructured conversation can also be pleasurable since many professors enjoy talking about teaching and do not do so as often as they would like. The consultation can also be structured around a student evaluation or a classroom visit by the consultant.

Visits in class can be a natural extension of consultation since they give the instructor and the consultant more to talk about. Unfortunately, most professors are not trained in classroom observation, and the correlation coefficient between the ratings done by different faculty raters after visits is $r = 0.26$ (Marsh, 1984), which is quite low. Despite this, peer visits are useful since the professor visiting the class is likely to provide some feedback, both positive and negative, that the students do not. Student evaluations are much more reliable than faculty evaluations, possibly because the students see the professor many more times than a professor visiting the classroom does. Although there are advantages to an unstructured procedure during visits (Elbow, 1986), correlation coefficients are likely to be higher if a structured procedure is followed. Acheson (1981) discusses one such procedure developed for college classrooms, while Andrews and Barnes (1990) discuss several of the highly structured instruments which are used for evaluating primary and secondary schools. If engineering colleges are ever forced to make assessments, the wealth of experience from primary and secondary schools should be used to show what does and does not work.

Administrative ratings are similar to peer ratings (Greenwood and Ramagli, 1980). An administrator often bases her or his ratings on informal information gathered from students. Administrators have one disadvantage compared to professors in visiting classrooms and in doing evaluations. Untenured professors in particular are likely to be intimidated by them. The advantage that department heads have is that it is part of their job to help young faculty improve their teaching, and many young faculty members report that such a person was the only professor with whom they had discussed teaching (Boice, 1991).

A systematic follow-up of alumni is quite appealing. Many professors argue that the alumni are older and hopefully wiser, have a feel for what is important in industry, and rate professors differently than students. Alumni follow-ups routinely result in very high agreement with ratings by current students (Canelos and Elliott, 1985; Centra, 1980). Since evaluations from

current students are cheaper, are easier, and result in a higher rate of usable returns, the college of engineering at Pennsylvania State University stopped doing alumni ratings of professors and switched to student ratings (Canelos and Elliott, 1985).

Discussions with engineers in industry are useful as part of a content evaluation. Students are perhaps least able to evaluate content, which is probably best done by a team of professors and engineers from industry. One advantage of ABET visits is that content is evaluated, albeit by only one person. In general, new professors who do not have industrial experience are likely to err on the side of being too abstract. Professors heavily involved in research are likely to put too much of their research in courses. Older professors who are doing neither research nor consulting may be presenting obsolete material.

Videotape has some application in course evaluations, particularly in considering some of the performance aspects of teaching (Centra, 1980). However, the presence of a video camera in the classroom can inhibit both professor and students. The result is a somewhat artificial class which will not be completely representative. Elbow (1986), who tried videotaping classes, was not sure that it was particularly helpful, and he noted that if a videotape is shown to a consultant, the professor should pick one that the professor is satisfied with. Our conclusion is that videotaping is probably worth doing once so that the professor can watch for annoying mannerisms.

Many critics of student evaluations claim that what should be analyzed is student learning or student achievement. As noted in Section 16.3.2, there is a positive correlation between student ratings and test scores. Although direct measurement of student learning to evaluate courses may be preferable, it is difficult (Centra, 1980; Davis and Alexander, 1976; Greenwood and Ramagli, 1980). One major difficulty is that students vary tremendously both within a class and from year to year. Should the evaluation be considered positive if many students score well on the test even though they may not have learned much new material? In other words, should it be the increase in knowledge or the total knowledge that counts? Should the learning of the better or the more poorly prepared students be counted differently? The better prepared students will probably score higher on tests but may learn less new material than students with poorer preparation.

Another problem with direct measurement of learning is that some type of standardized test must be used. Instructor-prepared tests can easily be written to cover what the instructor thinks the students know. This biasing of the test may well be unintentional. An alternative problem is that the instructor may teach to the test if he or she knows what is covered on the exam.

Measures of learning must include all levels of the taxonomies which are important in the course objectives. In the cognitive domain it is easiest to test at the three lowest levels, but certainly analysis, synthesis and evaluation are important in engineering education. The affective domain also needs to be included. Most professors and students would agree that a course in which students learn the material but hate it is not a good course.

Despite these problems with the direct use of student learning for the evaluation of teaching, it should be used to supplement other evaluation methods. In particular, student learning should be used for course improvement. Tests should be analyzed first for discrimination (see Chapter 11.2.2) and then to see if there are topics which students are not learning. If there are, then extra time or a different teaching strategy is needed. Once the problem areas have been pinpointed, the problems and possible solutions should be discussed with another professor.

Often professors try to teach too much material, and the easiest way to increase student learning is to cover less material but do a better job with that material.

Classroom observations or "classroom research" can also be used to determine what the students in a classroom are learning (Cross, 1991). One assessment technique is "minute papers." Toward the end of the class ask the students: (1) What is the most important thing you learned today? or (2) What questions do you still have? Not only do minute papers require the students to be active and construct their own knowledge, but they also provide useful feedback to the instructor. A perusal of the students' responses may show where your message is not getting across.

16.5. CHAPTER COMMENTS

The style in this chapter differs from that of previous chapters in that we have tried to cite all our facts, and in some paragraphs almost every sentence has a reference citation. This was done because of the controversial nature of evaluations of teaching. We wanted to be sure that our facts were backed by the research literature on evaluating teaching, and that skeptical readers could check our sources.

We are in favor of student evaluations and other methods of evaluating teaching since we believe that they help improve the teaching of undergraduates. Naturally, all these methods could be improved. However, there does not seem to be a justification for not evaluating teaching just because improvements are needed. There is clearly enough empirical evidence to show that student evaluations can separate good teachers from poor teachers. On the other hand, there is also evidence that student ratings are not a fine instrument and, for example, we cannot say that someone who ranks third out of twenty faculty is necessarily better than someone who ranks fourth.

16.6. SUMMARY AND OBJECTIVES

After reading this chapter, you should be able to:

- Discuss the advantages and disadvantages of formative and summative evaluations.
- Explain the various uses of teacher evaluations and discuss the controversies surrounding them.
- Discuss the various types of student ratings and how they should be administered.
- Discuss the reliability of student ratings and contrast it to the reliability of other evaluation methods.
- Discuss the validity of student ratings. Defend a position pro or con that student ratings are valid.
- Delineate the extraneous variables which affect student ratings and outline a procedure

to minimize the effects of these variables.

• On the basis of your personality determine the type of courses in which you are most likely to do a good or a poor teaching job.

• Discuss other evaluation procedures and how they can complement student ratings to help improve teaching.

HOMEWORK

1 Informally discuss your teaching with a colleague.
2 Ask a master teacher if you can visit her or his class. Make arrangements to do so and then discuss teaching with the master teacher.
3 Develop a simple formative evaluation instrument to use in your classes.
4 Obtain a copy of your university's summative form for student evaluations. Evaluate the evaluation form. Is it adequate? If not, how could it be improved?

REFERENCES

Abbott, R. D., Wulff, D. H., Nyquist, J. D., Ropp, V. A., and Hess, C. W., "Satisfaction with processes of collecting student opinions about instruction: The student perspective," *J. Educ. Psychol.,* 82, 201 (1990).

Acheson, K. A., "Classroom observation techniques," Idea Paper No. 4, Center for Faculty Evaluation and Development, Kansas State University, Manhattan, KS, 1981.

Andrews, T. E. and Barnes, S., "Assessment of Teaching," in Houston, W.R., Haberman, M., and Sikula, J. (Eds.), *Handbook of Research on Teacher Education,* MacMillan, New York, chap. 32, 1990.

Aubrecht, J. D., "Are student ratings of teacher effectiveness valid?" Idea Paper No. 2, Center for Faculty Evaluation and Development, Kansas State University, Manhattan, KS, 1979.

Aubrecht, J. D., "Reliability, validity, and generalizability of student rantings of instruction," Idea Paper No. 6, Center for Faculty Evaluation and Development, Kansas State University, Manhattan, KS, 1981.

Boice, R., "New faculty as teachers," *J. Higher Educ.,* 62, 150 (March/April, 1991).

Canelos, J. J., and Elliott, C. A., "Further investigations of teaching and course effectiveness evaluation. An ongoing project at Penn State Engineering," *Proceedings ASEE/IEEE Frontiers in Education Conference*, IEEE, New York, 77, 1985.

Centra, J. A., "The how and why of evaluating teaching," *Eng. Educ.,* 205 (Dec. 1980).

Cohen, P. A., Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings," *Res. Higher Educ,* 13, 321 (1980).

Cohen, P. A., "Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies," *Rev. Educ. Res.*, 51, 281 (1981).

Cook, D. I., "Write your own questionnaire," *Eng. Educ.*, 353 (Jan. 1975).

Cross, K. P., "Effective college teaching," *ASEE Prism*, 27 (Oct. 1991).

Davis, R. H. and Alexander, L. T., "Evaluating instruction," *Guides for the Improvement of Instruction in Higher Education, No. 3*, Michigan State University, East Lansing, MI, 1976.

Eble, K. E., *The Craft of Teaching*, 2nd ed., Jossey-Bass, San Francisco, 1988.
Elbow, P., *Embracing Contraries: Explorations in Learning and Teaching*, Oxford University Press, New York, 1986.
Fowler, W. T., "Improved feedback to the lecturer," *Eng. Educ.*, 250 (Dec. 1978).
Greenwood, G. E. and Ramagli, H. J., "Alternatives to student ratings of college teaching," *J. Higher Educ*, 51, 673 (Nov./Dec. 1980).
Higgins, R. C., Jenkins, D. L., and Lewis, R. P., Total quality management in the classroom: Listen to your customers," *Eng. Educ.*, 12 (Jan./Feb. 1991).
Jakubowski, G. S., "What students think about their teachers," *Eng. Educ.*, 372 (Feb. 1982).
Janners, M. Y., and Tampas, P. M., "Developing policies and procedures for evaluating instruction," *Eng. Educ.*, 675 (April 1986).
Johnson, G. R., *Taking Teaching Seriously: A Faculty Handbook*, Texas A&M University Center for Teaching Excellence, College Station, TX, 1988.
Koushki, P. A., and Kuhn, H. A. J., "How reliable are student evaluations of teachers?" *Eng. Educ.*, 362 (Feb. 1982).
Kuriger, W. L., "Some statistics regarding student-faculty evaluations," *Eng. Educ.*, 211 (Nov. 1978).
Lindenlaub, J. C., and Oreovicz, F. S., "A course evaluation guide to improve instruction," *Eng. Educ.*, 356 (Feb. 1982).
Lowman, J., *Mastering the Techniques of Teaching*, Jossey-Bass, San Francisco, 1985.
McKeachie, W. J., *Teaching Tips: A Guidebook for the Beginning College Teacher*, 8th ed., D.C. Heath, Lexington, MA, 1986.
McKeachie, W. J., "Research on college teaching: The historical background," *J. Educ Psychol.*, 82, 189 (Spring 1990).
March, H. W., "Student's evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility," *J. Educ Psychol., 76*, 707 (1984).
Milligan, M. W., "Evaluating teaching effectiveness for administrative decisions," *Eng. Educ.*, 374 (Feb. 1982).
Murray, H. G., Rushton, J. P. and Paunonen, S. V., "Teacher personality trait and student instructional ratings in six types of university courses," *J. Educ. Psychol.*, 82, 250 (1990).
Sherman, B. R. and Blackburn, R. T., "Personal characteristics and teaching effectiveness of college faculty," *J. Educ. Psychol.*, 67, 124 (1975).
Wilson, R. C., "Teaching effectiveness: Its measurement," *Eng. Educ.*, 550 (March 1972).