

Operating the Faculty Evaluation System: Issues in Designing and Using Student Rating Forms

LEGAL ISSUES

Easily the largest and most visible component of a faculty evaluation system is the student rating form and its computerized output. Over the years, for good or ill, student ratings have come to be the single most heavily weighted component of faculty evaluation systems. In response, many institutions have developed their own student rating forms, generally designed either by faculty, students, administrators, or a committee made up of some combination of these. Experience has shown that in the majority of cases these student rating forms have not been constructed in accordance with professional psychometric principles and standards. Thus, the forms may be of inadequate, or at least indeterminate, reliability and/or validity. The use of these forms may pose a legal liability for the institution.

In the general area of faculty evaluation the courts have tended to accept faculty-determined criteria and standards of professional performance, focusing instead on the consistent application of contractual conditions or procedures (Kaplin, 1978). Centra (1993) points out that, from a legal perspective, institutions must take care to comply with faculty evaluation procedures specified in the contract or faculty handbook. In addition, an institution may not use criteria or processes that create a discriminatory assessment based on ethnic background, race, or gender (Braskamp & Ory, 1994). Thus, the issues of *what* to measure and *how* to measure it have tended to be left to faculty to determine. Primarily, institutions must have procedures in place that do not violate basic civil rights and, in the event of an audit, must be able to provide evidence that their procedures have been applied correctly and consistently. The legal ramifications surrounding the

design and use of student rating forms in a faculty evaluation system must be considered within the context of these principles.

Carr & Padgett (1992) reviewed the legal liability in all 50 states concerning the design and use of student rating forms. They concluded that the primary legal issue revolves around statistical validity and reliability of such forms. Legal liability is posed when unreliable and invalid student rating forms are used in support of personnel decisions. The legal liability of the institution includes discrimination, defamation, and violations of individual privacy and academic freedom.

State laws provide an additional environmental element creating possible legal liability. Of particular concern are the open records laws. Unless exceptions are made, these laws provide that employment records such as salary and performance evaluations of public employees are public information. Some colleges and universities publish the results of student ratings of faculty. Some do not. Under open records laws, ratings would have to be made public for a single faculty member or for all faculty members if an individual or group requested this information. In this litigious society liability for publishing student evaluations of faculty members based on forms of dubious statistical validity seems rather apparent. (Carr & Padgett, 1992, p. 69)

An examination of the states' open records laws indicates that all states have laws allowing public access to government records; however, various states provide certain ex-

ceptions or limitations relative to access to faculty personnel records.

All Personnel Records Public—No Exceptions

The following states provide for no exception to state statutes pertaining to the release of public employee records, including faculty personnel records:

- Alaska
- Arizona
- Arkansas
- Georgia
- Massachusetts
- Maine
- Minnesota
- Missouri
- Montana
- North Carolina
- North Dakota
- Ohio
- Tennessee
- Texas
- Utah
- Virginia
- Wisconsin

All Personnel Records Public—General Exceptions

The following states grant only general exceptions pertaining to the release of faculty personnel records. These exceptions pertain to withholding information that would invade a person's privacy but do not specify what this information may be.

- Washington, DC
- Kentucky
- Louisiana
- Nevada
- Pennsylvania
- Washington
- West Virginia

Specific Exception for Faculty

The following states specifically exclude faculty personnel records, including student ratings, from release to the public:

- Florida
- New Mexico
- Oregon
- South Carolina

Personnel Records Not Public

The following states specifically prohibit the release of personnel records including faculty evaluation results:

- Alabama
- California
- Colorado
- Connecticut
- Delaware
- Hawaii
- Iowa
- Idaho
- Illinois
- Indiana
- Kansas
- Maryland
- Michigan
- Mississippi

- Nebraska
- New Hampshire
- New Jersey
- New York
- Oklahoma
- Rhode Island
- South Dakota
- Vermont
- Wyoming

Institutions in states that have no or only general exceptions to public records laws may find themselves at greater risk for possible legal action revolving around the publication of faculty records that include data from unreliable and invalid student rating forms. In those states where public disclosure of faculty records is less likely owing to specific exceptions, the risk of legal liability still exists if faculty personnel decisions are challenged in court.

In terms of the design or adoption of student rating forms, the key issue is to establish their reliability and validity prior to their use in a faculty evaluation system. The adoption of commercially available student rating forms such as those reviewed in Chapter 15 ensures that the forms are reliable and valid, since they have undergone rigorous psychometric procedures in their development. It is generally recommended that an institution consider the adoption of a commercially available student rating form before embarking on the task of building one from scratch. However, if the institution chooses to design and use its own student rating form, the administration must be willing to commit to the yearlong rigorous psychometric process involved producing a reliable and valid instrument. To fail to do so places the institution at unnecessary legal risk.

Civil rights legislation has provided a legal environment in which employment tests and performance evaluations must in most cases be statistically reliable and valid. That is, there must be statistical documentation that they are reliable measures, are job related and predict job performance. Otherwise, the employer's legal defense is likely to be inadequate in cases of discrimination. (Carr & Padgett, 1992, p. 68)

■ DESIGNING A VALID AND RELIABLE STUDENT RATING FORM

As frequently noted in previous chapters the design and development of a valid, reliable form intended to measure the teaching performance of an instructor and/or the perceived effectiveness of a course, is a technical task requiring professional expertise in statistics and psychological measurement. It is a common fallacy among educators that all that is required to develop a questionnaire is to sit down

and write a set of questions. It should be noted that what is being constructed is an instrument designed to measure psychological phenomena (i.e., perceptions, opinions, reactions, etc.). As such its development must be held to the same high professional standards of accuracy, reliability, and validity as any other scientific measurement tool or instrument. In fact, since non-physical (psychological) phenomena are being measured, the task of designing the measurement tool is that much more complex and difficult to accomplish. To assume that all one has to do to construct a psychometric instrument is to write a set of questions is analogous to assuming that all one has to do to construct an automobile is to get a motor and attach some wheels to it. Both tasks are easier said than done. And both tasks require not only precision and skill but also an understanding of the underlying principles and science involved.

As noted earlier, many student rating forms in use are homemade—constructed by faculty, students, administrators, and/or committees who lacked questionnaire design and scaling expertise. Such instruments typically are not subjected to reliability and validity studies and, therefore, may be easily influenced by extraneous variables in the classroom such as time of day, class size, and the instructor's personality.

In order for student ratings to be considered an integral part of a comprehensive instructional evaluation system, they must be reliable and valid. As a practical matter, certain issues must be considered when designing and/or selecting a student rating form.

Student Ratings Versus Student Evaluations of Teaching

One of the core issues that must be addressed in considering the process of designing a valid and reliable instrument to be completed by students as part of the assessment of faculty teaching performance is the differentiation between student ratings and student evaluations. As noted in the introduction, the term *evaluation* is defined as: The process of interpreting a measurement (or aggregate of measurements) by means of a specific value construct to determine the degree to which the measurement(s) represent a desirable condition. Considering this definition of evaluation, and examining the role of student input into the evaluation of teaching performance, we can take one of two directions: 1) Construct a form that asks students to make judgments on the desirability of certain instructor performances or characteristics (i.e., evaluate teaching and the teacher), or, 2) construct a form that measures the impact of the instructor's performance on the students and from which a judgment on the desirability of the per-

formance (evaluation) will be made later by the systematic application of a specific, predetermined value construct.

Taking the first approach results in constructing student *evaluation* forms. On these forms students are asked such questions as "How would you rate this instructor compared to all the other teachers you have had?" or, "Overall, how would you rate this teacher?" These questions ask the student to classify the faculty member in such ways as "among the top 10%," "exceptional," "unsatisfactory," or "among the worst." The problem with this approach is that the judgment about whether the faculty member's performance represents a desirable or undesirable condition is being made on the basis of the student's individual value construct. Each student's individual value construct may not only be unknown to either the faculty member or the administrator reading the results, but may differ from student to student. It is for this reason that student evaluation forms (that is, forms designed only to gather student evaluative judgments concerning faculty performance) may be considered inherently flawed. The reason for this is that no matter how technically correct the items may be in terms of their psychometric characteristics, there is no assurance that the value construct being applied by the student to interpret their observations and experiences (and thus make evaluative judgments) either matches that of the faculty or is even consistent from student to student. This flaw in many of the forms used in higher education generates the common conception that student evaluations are merely a popularity contest.

In contrast to student evaluation forms, student *rating* forms take the approach of measuring student perceptions and reactions rather than asking for an evaluation of some aspect of teaching. These forms measure those psychological phenomena that the student is best qualified to provide—their own reactions. Well-designed student rating forms include such items as "The objectives of the course were clear," "The audiovisuals used in class helped me to learn the material," and "The instructor's speaking style made it difficult for me to understand what he/she was saying." Students are then provided a "Strongly Agree—Strongly Disagree" or similar scale with which to respond to each item. In this way, measures of the students' perceptions and reactions to certain aspects of the instructor's performance are gathered. The evaluation of the instructor's performance is then made by someone else using an agreed-upon value construct to interpret the students' reactions as being indicative of either a desirable or undesirable teaching performance. As noted earlier, the use of a previously agreed-upon value construct in the evaluation process is the essence of *controlled subjectivity*. And, it

is controlled subjectivity that makes possible a fair and consistent faculty evaluation system.

Unfortunately, the terms *student rating* and *student evaluation* have come to be used interchangeably in higher education, and even in the professional literature. Not all so-called student evaluation forms truly call for the student to make an evaluative judgment, but are, in fact, student rating forms (although they may contain one or two items calling for an evaluative conclusion). On the other hand, not all student rating forms are designed to simply measure student perceptions and reactions, but include many items that call for students to come to an evaluative conclusion. It is likely that it is this mixture of terms and approaches (i.e., measurement of student perceptions and reactions versus requests for students to draw evaluative conclusions) that has led to much of the confusion and continuing concern over the use of student input in the evaluation of faculty performance.

In any case, as has been described in greater detail elsewhere, the evaluation of teaching performance must be based on multiple sources of data, not just student rating data which, of necessity, can only provide part of the picture of the complex (meta-) professional performance we call teaching.

Validity and Reliability of the Student Rating Form

The validity of student ratings is an issue of continuing discussion and disagreement, perhaps due in part to the confusion between the concepts of student ratings and student evaluations (the underlying reasons for this continuing confusion are explored in Chapter 12). Basically, the source of this disagreement is the fact that a universally accepted definition or description of the characteristics of an excellent teacher does not exist. While the student rating literature provides an indication of the aspects of faculty classroom performance that impact learning, there is still no precise formula for being a good teacher to which everyone subscribes.

Although from a professional psychometric perspective the validity of a given student rating form may be calculated, determination of validity is predicated on some underpinning construct or assumptions of what should be measured. Should a student rating form measure only those things that affect immediate learning outcomes? Should a student rating form measure some longer-term impact on students' perspectives, motivations, or values? Should a student rating form measure the degree to which a faculty member's style of dress or speech offends students? Faculty will answer these and many similar questions differently depending on what they consider to be

relevant and important (of value). Thus, from a practical perspective, the real problem in determining the validity of any student rating form is to determine a priori what aspects of teaching performance should be measured. There is no fixed answer to this problem. Although there is strong evidence that teacher preparation and organization positively affects student learning (Pascarella, 2001), different faculty value systems may dictate that different aspects of faculty performance (beyond those that affect immediate learning outcomes) be measured. For example, in certain faith-based institutions a faculty member's teaching performance may be required to include an expression of faith in the style of delivery and the context in which the content is presented.

In addition to the issue of the validity of student rating forms, there is the issue of reliability. From a practical perspective, a student rating form is reliable if it accurately and consistently measures students' perceptions and reactions to previously agreed-upon faculty teaching performances and characteristics. If the form does not measure those aspects of teaching performance that the faculty feel are relevant and important, it doesn't matter how accurate and consistent its measurements are. In fact, experience has shown that faculty prefer a form that they perceive as valid even if the evidence shows that the form is not reliable. Many institutions fall into the trap of using a student rating form that appears to be measuring the "right things," without being concerned whether the form is in fact producing consistent and accurate measures of those right things.

From a strict, professional psychometric perspective, an instrument must be reliable before it can be determined whether it is valid. That is, the instrument must be measuring something with accuracy and consistency before it can be confirmed that it is measuring what we want to measure. Unfortunately, however, in common practice institutions accept a student rating form as being valid in the sense that it asks questions that the faculty and/or the administration think should be asked, without checking to make sure that the form is reliable first. Such institutions may be at risk (see discussion on legal issues).

Designing a Student Rating Form

The following is a brief discussion and overview of the procedures required in building a valid and reliable student rating form. This section is not intended to serve as a technical reference on the issue of psychometrics. Rather, what follows is a summary, with examples, of the various phases an institution must go through to design and develop its own valid and reliable student rating form. It is assumed in this discussion that the institution will consult

with appropriate professionals in psychometrics when carrying out the following procedures.

Phase I: Determine the issues to be measured by the form. While the student rating literature indicates that the broad dimensions of teacher preparation and organization, clarity of communication, and rapport with students are among those that tend to impact student learning, it is important at this point to gain consensus among the faculty and administration on the major issues they feel should be considered. For example, if there is a concern that sexism or racism not be expressed or demonstrated in any way by the teacher, then this should be made clear. It is critical that the individuals involved in this discussion do not fall into the trap of beginning the process by simply sitting down and writing the items for the form. This discussion should steer away from specific questions and focus on the underlying issues and values on which the questions may be based.

The following strategy has been employed successfully in carrying out this step:

- Ask the faculty to develop a list of one-, two-, or three-word descriptors of an *excellent* teacher and another list of one-, two-, or three-word descriptors of a *poor* teacher.
- Ask the students to develop the same two lists.
- Examine the faculty-generated and student-generated lists of terms used to describe excellent and poor teachers and find those descriptors used by the faculty and students to describe each type of teacher.
- Use the list of common descriptors (those used by both faculty and students) for excellent and poor teachers to develop items for the student rating form.

Phase II: Write or select the items. Once the issues or characteristics to be measured are determined, the process of either writing new items or selecting them from existing catalogs (such as the one in Chapter 14) may proceed. As much as possible, the items should use the common list of descriptors determined in Phase I.

- Each item should be written to measure only one perception, reaction, etc.
- For each issue to be measured be certain to write and/or select at least six to eight items.

Phase III: Develop/select appropriate response scales. A key element in building a valid and reliable student rating form is the scale that is used to elicit student responses. It is critical that the response scale used for each item meets certain logical and technical requirements, including:

1) The response scale must be *appropriate*. Response choices must logically follow from the item/question. For example, the item "Was the instructor organized?" logically calls for a "yes" or "no" response. To ask such a question and then give the student a response scale such as Excellent–Poor or Strongly Agree–Strongly Disagree will result in unreliable data.

2) The response scale must be *parallel*. That is, the scale must not mix two or more types of responses. The following is an example of a *nonparallel* scale:

- Very Good
- Above Average
- Satisfactory
- Unsatisfactory

Very Good fits with a Very Good–Very Poor scale; Above Average fits with a scale that provides indications of performance above, at, or below Average; and Satisfactory and Unsatisfactory belong to a separate satisfaction scale. A nonparallel scale compromises the reliability of the student ratings.

3) The response scale must be *balanced*. There must be an equal number of positive and negative choices. An *unbalanced* scale is one that has either more positive or more negative choices. Obviously if you have more opportunities to make a positive response than a negative one, the responses will tend to be skewed in that direction. The converse applies if there are more negative responses than positive. Also, it is generally recommended that odd-numbered scales in which the middle response is "Neutral" or "Don't Know" *not* be used unless the middle position has a specific meaning relative to the item. Otherwise some students will use the middle response simply to avoid answering the item. The following are examples of a balanced, parallel scale and a scale that is unbalanced and nonparallel:

Balanced Parallel Scale	Unbalanced and Nonparallel Scale
(+) Strongly Agree	(+) Excellent
(+) Agree	(+) Outstanding
(+) Disagree	(+) Satisfactory
(+) Strongly Disagree	(+) Unsatisfactory

4) Every point on the scale must be defined. Do *not* use a scale where only the end points are defined such as "Low 1 2 3 4 5 High." With such a scale you can never be sure what the student is thinking when they

pick a "2," for example. Such scales tend to compromise the reliability, and perhaps even the validity, of the response data.

- 5) Make certain to use abbreviations to denote points on the response scale, *not* numbers or the letters A, B, C, D, E. For example, if your response scale is Strongly Agree–Strongly Disagree, use the abbreviations SA for Strongly Agree, A for Agree, D for Disagree, and SD for Strongly Disagree. Do not use 4 for Strongly Agree, 3 for Agree, and so forth. Also, do not use A for Strongly Agree, B for Agree, C for Disagree, and D for Strongly Disagree. Again, when either numbers or letters are used instead of the abbreviations for the responses, you can never be certain what the students are thinking when they respond. The letters A, B, C, and D can easily be interpreted as grades rather than varying degrees of agreement. The numbers 4, 3, 2, and 1 can mean different things to different people, including being interpreted as ranking system. Therefore it is critical that the response choices on the student rating form either use the full words describing the choice points or an appropriate abbreviation of those words. •

Phase IV: Conduct field trials to gather the data needed for subsequent validity and reliability determination. At this point you may be testing 50, 60, or even 100 items. Separate subsets of these items should be placed on different forms and administered to all kinds of classes. Once you have at least 200 responses for each item (that is, each item will have appeared on one form or another so that at least 200 students have had the opportunity to respond to it) you are ready to move to the next step.

Phase V: Conduct a factor analytic study. It is critical to determine the reliability and validity of the form and the degree to which the items reliably measure each issue or characteristic of interest. Nationally standardized forms, such as those reviewed in Chapter 15, should be used as the vehicle for developing your own form: Insert the items developed for your form in the optional items section of the commercial form and then conduct factor analytic studies on the responses to the commercial rating form items and the items written for your form. In this way it is possible to compare the results of the analysis of your items with the analyses of the items on the commercial form. Although the factor analytic study must be conducted by individuals proficient in statistics and psychometrics, the outcome of such an analysis should produce at least four items for each issue of interest. In order to adequately measure a specific perception of or reaction to an

issue of interest, at least four items measuring that perception or reaction must be used. From a technical perspective the factor loadings of these items on the dimension being measured should be .400 or higher.

Phase VI: Develop subscales based on the result of the factor analysis. The form should produce separate scores for each of the dimensions or characteristics being measured by the form. Such subscales provide more detailed diagnostic information for instructor use in improving or enhancing their performance and fairer comparisons across time and across different courses.

Phase VII: Refine the form based on the results of the factor analytic study to ensure reliability of each subscale and the form as a whole. Ideally, the reliability values should be .60 or higher.

Phase VIII: Establish norms. Since no specific, agreed-upon set of criteria for what constitutes an excellent teacher exists, we must use a norm-referenced model of evaluation. That is, we cannot simply say that someone is an excellent or poor teacher based on a specific rating value derived from the student rating form. It is important to develop a way to interpret the ratings relative to an appropriate cohort group. This part of the process is the most time consuming, since data on the form must be gathered for at least three years before any kind of appropriate normative interpretation may be made. Data from all classes should be accumulated without replacement into a normative database that can then be used to properly interpret the student ratings of an individual faculty member. Thus, it will become possible to interpret the student ratings for an individual in terms of where the faculty member stands relative to others in the norm group. The commercially available forms reviewed in Chapter 15 all make use of norms—developed over decades of use across a broad spectrum of courses and course types—to provide appropriate interpretations of an individual's ratings.

Phase IX: Organize the items. The final step in developing your student rating form is to decide how the items are to be grouped and labeled in the final design of the form: Determine how the items should be organized for easy reading and answering and how and where the responses should be recorded. If there is a logical or chronological flow to the items, then their organization on the form should reflect that. If there are only a few negative items, then one or two should appear very early in the instrument to avoid positive response set mistakes (i.e., before students get in the habit of thinking the response to the extreme right is always the most positive). It is advisable to have a few negatively stated items in the instrument, but only if they can be stated negatively in a logical manner.

At this point it should be clear that constructing a valid and reliable student rating form is a time-consuming, technical process that requires various studies and analyses. As stated before, it is recommended that experts in psychometrics and statistics be consulted in undertaking such a project. A good example of a study that demonstrates the technicalities involved in constructing a valid and reliable student rating form is "Differential Relationship of Student, Instructor, and Course Characteristics to General and Specific Items on a Course Evaluation Questionnaire" (Aleamoni & Thomas, 1980).

Phase X: Implement the Student Rating System. Once the student rating form has been completely developed and has been shown to be valid and reliable, it can be placed into operation within the comprehensive faculty evaluation system. The overall usefulness of the student rating form can be maximized, however, if it is constructed so that it has not only the standard, validated, and reliable items, but also room for additional optional items that may be unique to either a department or an individual faculty member. Forms constructed in such a manner allow the instructor to select supplementary (or more diagnostic) items from a catalog. Each commercial student rating form reviewed in Chapter 15 is designed in this manner. For example, the Course/Instructor Evaluation Questionnaire (CIEQ) provides the standard 21 items (from which the five subscale scores are derived and for which normative data are provided), and two optional 21-item response sections. The optional item response sections allow instructors to select up to 42 additional items from the 373-item CIEQ optional item catalog.

A more sophisticated approach to maximizing the utility of student rating forms is to develop multiple standard forms. This requires the institution to possess a great deal of in-house expertise in psychological measurement and statistics and be willing to undertake the major research and development projects involved. Under this approach optional items and optional item sections are generally not available to the individual instructor. Instead the instructor is given a choice among forms that more closely match specific teaching/learning environments. For example, the University of Washington Instructional Assessment System provides six valid and reliable forms to their faculty, one for small lecture/discussion courses, one for large lecture courses, one for seminar courses, one for problem-solving courses, one for skills-oriented or practicum courses, and one for quiz sections (Abrami & Murphy, 1980).

■ ADMINISTERING STUDENT RATING SYSTEMS

After an appropriate instrument has been developed or selected, administrative procedures need to be established. If possible the responsibility for managing and directing a campus-wide program of distributing, administering, and collecting student ratings should be given to a unit such as a teaching and learning center, a professional development center, a testing and evaluation center, or an office of institutional research. The responsibility for operating the system should be given to the student government or distributed to either individual college deans or department chairs. As a last option, the responsibility for operating the system may be assumed by the chief academic officer of the institution, since the danger of having the program perceived as a "watchdog" program of the administration is increased by doing so. Ideally, the unit that operates the student rating system should be seen, legitimately, as one dedicated to providing services to the faculty.

In keeping with the concept of designing and operating the faculty evaluation system to serve formative and summative purposes, several guidelines related to student rating systems may be stated. First, it must be recognized that the value of any student rating system relies on the confidence the students have that their input will cause them no harm, and have some effect on the instructor (Arreola, 1987). In actual practice, operating a student rating system requires a careful balancing of the needs and concerns of the students, faculty, and administration. These needs and concerns are sometimes antithetical to one another. This can often place the person in charge of the office or agency running the student rating system in a very difficult and tricky situation.

On the one hand, faculty may be fearful and distrustful of the administration. Faculty go through a predictable set of stages in resisting or attempting to escape from a faculty evaluation system. On the other hand, students may be fearful of retribution by the faculty if they give negative ratings, and they might not believe the faculty or the administration will pay attention to what they say anyway. Finally, the administration may desperately want any kind of quantitative or hard data on which to base difficult personnel decisions. Long experience with these circumstances leads to the following practical guidelines pertaining to running a student rating system:

- 1) If at all possible, do not locate a student rating coordination or processing agency in the office of a dean, vice president, or other major administrator. Such placement only reinforces the idea that the student

rating system and the faculty evaluation office are simply watchdog agencies of the administration.

- 2) Try to locate the student rating form processing office in a test-scoring center, computer center, media center, or, ideally, in a faculty professional enrichment center.
- 3) Arrange the processing schedule so that the completed analyses of the student rating forms are not available until after final grades have been reported.
- 4) Conduct a program to maintain the credibility of the students in the student rating system. Include regular contacts with student government, appoint student representative(s) to the Committee charged with developing the comprehensive faculty evaluation system. Submit articles to the student newspaper describing the importance and use of student ratings at least once each term. A constant communication campaign with the students must be maintained which informs them that the faculty member will not see the student rating results until after grades have been reported and that the student ratings are taken seriously by the faculty and the administration (Arreola, 1983). Without such a campaign, the student rating system will experience serious problems, including refusal of the students to complete forms, completing forms by simply marking the same response for all items, and covering the forms with various types of graffiti.
- 5) Make it clear that the student rating form processing office is a service to the faculty and not the administration or the student government. Do not automatically send results of the ratings to the administration or the student newspaper, even though written permission has been given by the instructor. Such actions will forever taint the credibility of the processing office. Require anyone wanting information concerning the student ratings of a faculty member to get them from that faculty member.
- 6) Make certain that the issue of the distribution of copies of the individual faculty student rating reports is a matter between the administration and the faculty or between the students and the faculty. Provide the faculty member with multiple copies of the student rating report form containing the analyses or other results; do not keep copies in the processing office. Make it physically difficult to recover or reconstitute a given faculty member's student rating report. The best approach is to maintain only raw data computer

files. Then, if a request is received by the processing office to provide copies of a particular faculty member's evaluation results, it can truthfully be said that the processing office has no copies but that the faculty member has several.

The issue of the confidentiality and distribution of individual faculty member's student rating reports should be a matter between the faculty member and the administration and not between the processing office and the administration. The processing office must not be perceived as an arm of a "big brother" administration.

The method of administering and gathering student responses can determine the quality of the resulting data. It is advisable to administer the instrument in a formalized manner in the classroom by providing a standard set of instructions and enough time to complete all the items. If the instrument is administered in an informal manner, without a standard set of instructions and a designated time to fill it out, the students tend not to take it seriously and possibly do not bother to turn it in. Furthermore, if the students are permitted to take the instruments home to fill them out and return them at the next class meeting, very few instruments will be returned.

The following practical procedures have been successfully used in managing large student rating systems. These procedures assume that the faculty member will not be the primary person administering the rating forms in class to the students.

- 1) Set up a log system for maintaining control of student rating form distribution and collection. This log should contain the name of the faculty member, the number of the course, and the enrollment. Such information is generally available from the registrar's office.
- 2) Prepare student rating packets with at least five more sheets than the official number enrolled in the course. Log in the actual number of sheets sent to the instructor.
- 3) In addition to the student rating forms, the packet should contain a standardized script to be read when administering the forms and a proctor identification form (PIF). Upon receipt of the packet, the faculty member should remove the PIF and indicate the name of the student in the class who has been selected to administer the rating forms. The faculty member must sign the PIF and return it separately to the processing office. The processing office should log in the date of the receipt of the form and the name of the student proctor.

- 4) After removing the PIF, the faculty member should give the student rating form package to the chosen student proctor. The student proctor removes and distributes the student rating forms and removes and reads a special form identified as the proctor administration form. It contains the standardized script for administering the student rating form. In addition to the standard information concerning the use of a #2 pencil to record student responses, the script should note that
 - The faculty member is not in the room.
 - The results of the rating will not be returned to the instructor until after final grades have been turned in.
 - The students' responses will be an important part of the information considered in improving the course or making promotion, tenure, retention, and merit pay decisions.
- 5) The proctor must sign the proctor administration form, certifying that the student rating forms were administered in accordance to the instructions, that the script was read as part of the administration, and that the faculty member was not in the room.
- 6) Finally, cross-checking items should be included on the student rating forms that ask such questions as, "Was the instructor in the room when this form was administered?" "Did the proctor read the administration directions out loud to the class?" and, "Do you have confidence that your responses will make a difference?"
- 7) After the forms are completed and returned to the proctor, they should all be placed in the envelope, along with the signed proctor administration form, and dropped in the campus mail to the processing office.
- 8) Upon receipt of the packet, the date of receipt should be logged in; the name on the proctor administration form should be checked against the name reported by the instructor on the proctor identification form. A count is made of the incoming completed student rating forms to make sure that the number does not exceed the official enrollment figure for the class. This latter step is designed to discourage stuffing of completed student rating packets by either students or, in certain instances, the faculty themselves.
- 9) Before machine processing, the student rating forms must be visually scanned for stray marks. Often the students doodle in the margins or simply cross out

an incorrect response rather than erase it. These types of marks must be erased before the sheets can be processed. Experience has shown that the student rating processing office would be well advised to buy electric erasers rather than using the eraser end of a pencil. The time and staff size required to carry out this necessary step is often much larger than would first be anticipated. This is especially true if the system is new and the students are not yet familiar with the rating forms.

- 10) Finally, log the date when the completed computer analysis and student rating forms were returned to the faculty member.

■ OPTIONS FOR ADMINISTERING RATING FORMS IN CLASS

The steps described in the previous section assume that a student proctor, selected from the class itself, administers the rating form. Other options, more or less desirable, are possible.

Self-Administered

In systems where instructors administer their own student rating forms, they should also read the standard set of instructions after the forms have been distributed and then remain in the front of the room until all forms have been completed. The instructor may then select a student from the class to gather the completed forms and deposit them in the campus mail. As before, the statement read should specify that the instructor will not see the results until after the term grades have been turned into the admissions and records office. The exception to this procedure is when the instructor has informed the students that their responses will be used in a formative way to improve the current course.

Student Government Administered

Another option is to have representatives of the student government association administer the student rating forms if the faculty and department or college administrators are willing or request them to do it. The students administering the instruments should also read a similar standard set of instructions and request that the instructor leave during the administration. The student organization could use the campus newspaper to announce when the instruments are going to be used and how they will be administered as a final cross-checking procedure.

Staff Administered

If an administrator decides to designate a staff member to administer the student rating form, then the same procedures should be followed as suggested above. This option should be avoided, however, unless there is no other way to ensure a common administration of the instruments. As noted earlier, faculty and students tend to feel threatened if they know that an administrator is controlling and directing the administration or processing of the instruments.

When the student rating form is administered, the students should have all necessary materials. Students should generally fill out the forms in their regular classroom near the beginning of a particular class session. Above all, the students must be left with the impression that their frank and honest comments are desired and not that this is their chance to get back at the instructor. If the students get the impression that the instructor is not really interested in their responses, they will not respond seriously. If the students feel the instructor is going to see their responses before final grades are in, they will respond more positively and write very few comments. This is especially true if the students are asked to identify themselves on the student rating form. If the form is administered immediately before, during, or after the final examination (or any other meaningful examination, homework, or term paper deadline), the students tend to respond in an inconsistent manner. If students are allowed to discuss the course and instructor while filling out the forms, then biases may enter into their ratings.

■ THE STUDENT RATING REPORT

One of the most important aspects of any program is the method of reporting the results. If the results are not reported in an appropriate, accurate, and timely manner, the usefulness of the student rating form and the system as a whole will be seriously compromised. When tabulating and summarizing item responses, the following procedures should be considered:

- The item responses should be given numerical values in order to calculate descriptive statistics, and the descriptive statistics should be reported. However, the response positions on the form should be labeled only with abbreviations and *not* the numerical values used in later computations.
- The results should be reported by item and subscale, if appropriate.
- The results should be summarized by class section, department, college, and so on.

Directionality of Numerical Scale

When items are presented with defined response scales, such as Strongly Agree, Agree, Disagree, and Strongly Disagree, they should be weighted to reflect direction and ideal response when the results are tabulated and summarized. For example, if this response scale were weighted 4, 3, 2, and 1, respectively, it would indicate that the item was positively stated with the ideal response being Strongly Agree. With such a weighting, it is possible to calculate a mean and standard deviation for each item for a given class of students. The mean value indicates their average rating and the standard deviation indicates how similar or dissimilar their responses were. With such a weighting scheme and the resulting means and standard deviations, the results can then be reported for items and subscales. The results can also be summarized and reported by class section, course, selected courses, courses within a department, courses within a college, and courses within the university. Such complete reporting schemes permit meaningful comparisons when necessary.

Distributing Rating Results—Voluntary Systems

An important aspect of any system of reporting student rating results is who will or should actually see the results. If the administration of the instruments is completely voluntary, then only the faculty members themselves should receive the results. As noted earlier, the processing office is ill advised to enter into an arrangement that places it between the faculty and the administration.

Distributing Rating Results—Mandatory Systems

If the administration of the student rating form is mandatory, as is the case in many systems, then every effort should be made by the processing office to remove itself from the responsibility of distributing rating results directly. However, if the processing office has no choice and the system must be designed to provide faculty with the option of releasing copies of their results to other interested parties, great effort must be taken to ensure that the instructor feels no pressure to release the results. In such a circumstance, a procedure must be implemented that requires a formal written release by the instructor. The processing office would be well advised to consult with the college or university attorney in developing the wording of such a release. Every effort must be made to make faculty members aware of who is to receive copies of their results and how frequently. Under no circumstances, however, should students' written comments be reported to administrators or student organizations, because those comments

tend to be susceptible to widely discrepant subjective interpretations by the reader.

Publishing Rating Results

If the faculty or administration has entered into an agreement with the campus student organization to publish the results of the student ratings, every effort should be made to ensure fair and accurate reporting of the results. Such results may be reported in a book divided by discipline or content area or published on a web site. Student-published books are usually promoted as course and instructor guides for prospective student enrollees as well as vehicles to encourage instructional improvement. Unfortunately, due to the zealousness of some student editors and the vendettas of others, such publications have tended to simply generate antagonistic relationships between the rated faculty and student editors.

Recently the Internet has become an outlet for student rating results. Student organizations may publish student ratings on a web site and, in a few instances, invite comments to be made via email. These anonymous email comments are then posted to the site. This approach to distributing faculty rating data poses new challenges, not the least of which is the issue of confidentiality. In a setting where it is illegal for faculty to publicly post student grades with identifying names and social security numbers, it is likely that the emerging practice of posting faculty ratings on the Internet will generate legal challenges in the future. In any case, care must be taken to respect the legal requirements of your state relative to the release of faculty personnel records as discussed at the beginning of this chapter.

■ FORMAT OF THE STUDENT RATING FORM COMPUTERIZED ANALYSIS OUTPUT

It has been assumed that virtually every operational student rating form system produces some form of computerized analysis output. This output may range from a simple frequency count for each response to each item up to a very sophisticated printout showing normative data of various sorts. Experience has shown that most faculty react to a sophisticated printout that contains page after page of indices, norms, graphs, and tables with something less than total enthusiasm. Although professional standards demand that the analysis of student rating data be accurate and statistically sound, it must be remembered that most faculty are not well versed in the intricacies of item analysis, measurement theory, or statistical analysis. Moreover, most faculty have no real desire to become well versed in these areas. Thus, for the computerized analysis of stu-

dent rating results to have the desired effect of providing useful information that may be acted upon by the faculty member for self-improvement, an effort must be made to make these computerized analyses user friendly. One way to do this is to provide a verbal summary sheet that translates the statistical information into general statements.

For example, if the student rating form has four response choices per item, it is clear that a standard deviation exceeding 1.00 would indicate wide disparity of responses to the item. In such a case, the mean value of the item would have little value other than to represent the numerical average of the responses. The computer program can be written to produce a statement such as the following:

On item 6, the standard deviation was 1.3. This can be interpreted to mean that there was CONSIDERABLE DISAGREEMENT among the students on this item and thus the average response value should NOT be interpreted as representing a consensus rating by the class.

Or, the program could present a statement such as

On item 18, the average response was 3.1 and the standard deviation was 0.4. This can be interpreted to mean that the students rated you as being MODERATELY HIGH on this item and there was a HIGH degree of consensus among the students in this rating.

Ideally, such statements should be printed as either the first or the last page in a computer analysis so they are easily found. Obviously, the program must be written to make the proper interpretations of various combinations of data relative to the appropriate norms. The point is that even though the processing office may produce reports which are statistically sophisticated and correct, one must never lose sight of the fact that for the analyses to be useful they must be understood and used by the faculty. As a practical guideline providing computerized, written interpretations of the statistical information listed in the printout is highly effective in helping to promote the effectiveness of the student rating information.

■ INTERPRETING AND USING STUDENT RATING RESULTS

Although we may want our analyses to be user friendly, they must still present data of sufficient clarity and detail to permit sophisticated interpretations. How accurately and meaningfully the results of student ratings are interpreted and used depends on the type of information pro-

vided to the participating faculty member and other interested parties. The research on student ratings has revealed a definite positive response bias, which needs to be addressed when interpreting and using the results. That is, if students are asked to respond to positively stated items using a 4-point scale of Agree Strongly, Agree, Disagree, and Disagree Strongly, the responses tend to be distributed as shown in Figure 13.1.

To someone unfamiliar with statistics, the midpoint of the 4-point scale (2.5) could easily be interpreted as average, and any rating higher than 2.5 could be interpreted as a positive rating. But, as Figure 13.1 shows, the easy interpretation that any rating of 2.5 or higher is good can be seen as being substantially in error. In fact, because the distribution of student ratings is skewed, the average or mean rating tends to fall closer to 3.0 than 2.5.

One effective way to report student-rating data is to present decile information based on the appropriate norm base. In this way, the bias of the student rating distribution can be taken into account. Figure 13.2 shows one way to represent student rating data in deciles.

The use of comparative (normative) data, such as the deciles shown in Figure 13.2, when reporting results can serve to counteract the positive response bias and result in a more accurate and meaningful interpretation of the ratings. For example, comparative data gathered on freshmen-level courses in the anthropology department allow the instructors to determine how they and their courses are perceived in relation to the rest of the courses in the department. When such comparative data are not available,

the instructor will be interpreting and using results in a void with very little substantiation for the resulting conclusions and actions taken.

Qualitative judgments can also be provided to the instructor by identifying course mean intervals in the comparative data, which can be defined as representing levels of excellence or needed improvement. For example, the comparative data for the freshmen courses in the anthropology department consisting of course means on student rating questionnaires could be divided into 10 equal portions. Each portion could be defined as representing a 10% interval of rated courses with a defined minimum and maximum course mean. These 10 intervals could then be defined as follows:

- *VP = Very Poor.* Any course mean falling in the lowest 10%, 20%, or 30% interval is defined as being of unacceptably low quality insofar as the elements being measured by the student rating form are concerned.
- *P = Poor (Improvement Required).* Any course mean falling in the 40% or 50% interval is defined as being of low quality (insofar as the elements being measured by the student rating form are concerned) and indicates that improvement is required.
- *G = Good.* Any course mean falling in the 60% or 70% interval is defined as being of solid professional quality (insofar as the elements being measured by the student rating form are concerned) and, although improvement is always desirable, no improvement is required.

Figure 13.1 Skewed Student Rating Response Curve

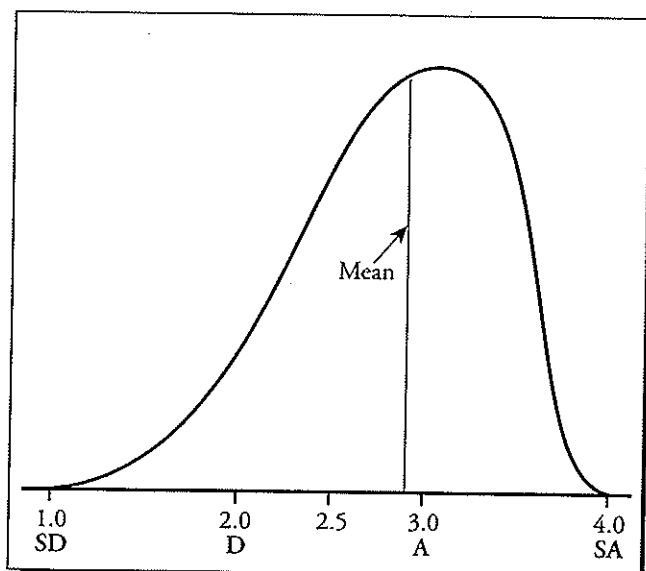
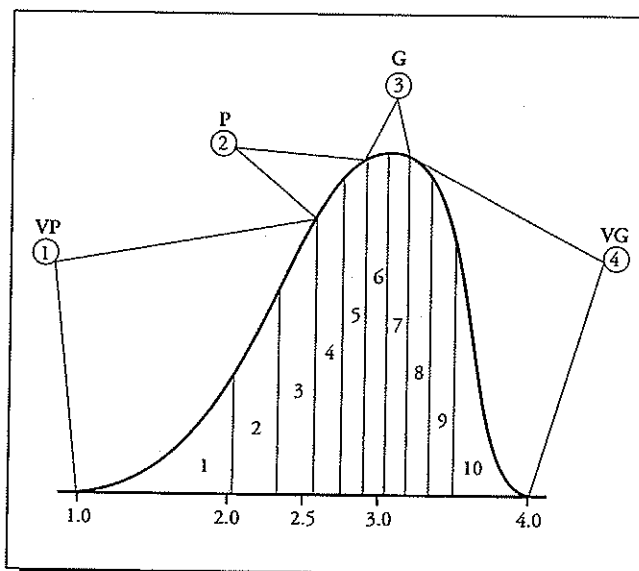


Figure 13.2 Decile Interpretation of Student Ratings



- *VG = Very Good.* Any course mean falling in the upper 80%, 90%, or 100% interval is defined as being of high quality (insofar as the elements being measured by the student rating form are concerned) surpassing standard professional performance.

This information could be provided to each participating instructor in a computerized format along with the appropriate interpretive materials. Or, as noted earlier, these values could be built into the computer program to produce appropriate written interpretive statements.

Further, the numerical values of 1, 2, 3, and 4 have been assigned to the labels VP, P, G, and VG, respectively. Thus, a raw student rating average of 3.2 may translate into either a decile of 8, a category rating of Very Good, or provide an appropriately norm-referenced summary qualitatively determined value of 4 since $VG = 4$. In computing the composite role rating (CRR) for teaching, the norm-referenced summary value of 4 in this example would be entered into the matrix in Figure 9.1 if this data were indicative of faculty performance in instructional delivery for instance. If student rating data from more than one course were being used in the matrix in Figure 9.1, then the norm-referenced value of 4 in this example would be averaged with the comparable data from the other courses.

Once established on a representative number of courses, the normative database should not change appreciably from year to year. Additional courses can be added to the normative database without significantly changing the distribution and comparative judgments. For some of the sources of invalidity identified as nontrivial by research studies, such as class level and required–elective status, comparative data stratified by course level and required–elective status will provide meaningful interpretations of the results. Aleamoni and Stevens (1986) include particular examples.

Once faculty members are provided with comparative data and interpretive materials, they are then ready to interpret their results. The comparative interpretations result from the normative data provided, and the subjective interpretations result from reflections on what took place in the classroom that could be related to the comparative interpretations. Using this procedure, in addition to a careful reading of the students' written comments (if available), each instructor should be able to generate diagnostic interpretations of instructional strengths and weaknesses in the course. If an instructor has results on two or more similar sections, then the ratings of one section may be compared to those of another section to determine what instructional behaviors may have led to the higher comparative ratings in one section.

If the faculty are not able to generate any diagnostic interpretations, they may need to talk with the department head, dean, or instructional development specialist about the results. This assumes, of course, that these individuals know how to interpret the results. Finding that this approach still does not adequately identify the source of instructional difficulty, the instructor may want to consider other procedures (e.g., the use of additional diagnostic optional items, classroom visitation, videotaping, etc.) in future evaluations.

After identifying their instructional strengths and weaknesses, instructors can use the information to plan an improvement strategy. In some instances, the strategy may simply require minor modifications in the course or teaching method. In other instances, the strategy may require a substantial commitment of time and resources on the part of the faculty member and the department. It has been through a process such as this that instructors have been able to use student ratings to identify instructional problems and then rectify them. Obviously, the success or failure of such a venture rests solely with the instructors and their willingness to gather and use the data provided.

If faculty members decide to submit copies of their student rating results to their department head or dean for rank, pay, and tenure considerations, then all of the appropriate interpretive materials should also be provided. Ideally, the student rating results should be interpreted by someone with expertise in the field of measurement and instructional development. Deans and department heads should be made aware of the necessity of using the comparative data to interpret the results rather than relying on the more subjective and highly unreliable written and oral comments of students. Interpretations of student rating results may also be carried out by peer review triads using several other assessments of instructional effectiveness such as self-evaluation, quality of student learning, and peer ratings of content (Aleamoni, 1987).

Student course/instructor rating data should never be used alone in evaluating instructional effectiveness for rank, pay, and tenure decisions, because such data are not completely diagnostic of all elements in the instructional domain. How such student rating data should be used in a comprehensive system of instructional evaluation and how much weight they should carry is something that should be determined at the department level.

■ FACULTY EVALUATION AND PROFESSIONAL ENRICHMENT

Experience has shown, time and again, that a faculty evaluation system implemented without reference or

connection to a professional enrichment program will generate greater amounts of anxiety and resistance among the faculty than if it is part of a larger faculty development or instructional improvement effort. Likewise, experience has also shown that faculty professional enrichment programs, operated in isolation or without reference to a faculty evaluation program, tend to attract mainly those faculty who need their services the least.

Ideally, a faculty evaluation system should be an integral part of a larger faculty evaluation/professional enrichment program. To achieve maximal benefit from these two programs, each element of the faculty evaluation system should have a corresponding and concomitant element in the professional enrichment program. Thus, if the faculty evaluation system is going to evaluate how well faculty teach courses or how frequently they publish scientific articles, there should be professional enrichment seminars, advanced workshops, and instructional materials available to assist them in learning how to teach better or how to write manuscripts that are more likely to be accepted for publication. In short, the faculty evaluation system should provide diagnostic information on the strengths and weaknesses a faculty member possesses and then follow up with programs or materials to aid the faculty member in enhancing strengths or overcoming weaknesses.

Again, as a practical matter, the computerized report derived from student ratings should include written comments that not only highlight the major areas of concern but also provide information on where to seek assistance. For example, a computer printout could include a statement such as:

On the TESTING AND GRADING section of the student rating form, the majority of the students (87%) indicated that your tests did not seem to relate well to the course objectives. The Office of Assessment and Professional Enrichment offers a seminar for faculty on Test Construction that may be of some interest. Call 555-1234 for information about the next seminar.

Thus, to be truly effective, a faculty evaluation program must be linked to, and work in concert with, a structured and supported professional enrichment program. Only in

this way will both programs stand a reasonable chance of achieving the common goals of improving instruction and enhancing faculty performance.

■ ONLINE STUDENT RATINGS

The administration and use of conventional paper-and-pencil student rating forms clearly involves cost and intrusion in class time. The advent of Internet technology, however, has made possible the use of online student rating forms which may be completed outside of class and which cost less to administer. Despite the continuing problem of low response rates to online student ratings, and even though the research concerning the utility, reliability, and validity, of online student ratings is just beginning to emerge, there appears to be a rush to adopt this technology. In fact, the popularity of online student rating systems has begun to grow to such an extent that more than 60 universities report their use (Clark, 2003). Online student rating systems are seen as having a number of advantages over paper-and-pencil systems including ease of administration, more complete data collection, longer, more thoughtful student responses, reduced processing time and costs, more accurate data collection and reporting, and more detailed, user-friendly reports (Hmielski & Champagne, 2000; Johnson, 2003). Theall (2000) has questioned these preliminary findings. The fact is that using technology to streamline data handling does nothing to address a fundamental lack of sufficient research concerning online teaching and learning, the quality of the underlying faculty evaluation systems making use of online student rating data, and the quality and appropriateness of the reporting and interpretation of such data (Franklin & Theall, 1989). In the larger context of considering the development of a comprehensive faculty evaluation system, it must be concluded that insufficient literature and experience exists within the higher education community to make definitive recommendations concerning online student ratings at this point. It is recommended, however, that institutions interested in this technology explore the literature that currently exists. See Sorensen and Johnson (2003) for an excellent starting point for entering this literature.

■ REFERENCES

- Abrami, P. C., & Murphy, V. (1980). *A catalogue of systems for student evaluation of instruction*. Montreal, Canada: McGill University Centre for Teaching and Learning.
- Aleamoni, L. M. (1987). Evaluating instructional effectiveness can be a rewarding experience. *Journal of Plant Disease*, 71(4), 377-379.
- Aleamoni, L. M., & Stevens, J. J. (1986). *Arizona course/instructor evaluation questionnaire (CIEQ): Results interpretation manual*. Tucson, AZ: Office of Instructional Research and Development, University of Arizona.
- Aleamoni, L. M., & Thomas, G. S. (1980). Differential relationships of student, instructor, and course characteristics to general and specific items on a course questionnaire. *Teaching of Psychology*, 7(4), 233-235.
- Arreola, R. A. (1983). Establishing successful faculty evaluation and development programs. In A. Smith (Ed.), *Evaluating faculty and staff, new directions for community colleges: No. 41* (pp. 83-90) San Francisco, CA: Jossey-Bass.
- Arreola, R. A. (1987). The role of student government in faculty evaluation. In L. M. Aleamoni (Ed.), *New Directions for Teaching and Learning, No. 31, Techniques for evaluating and improving instruction* (pp. 39-46). San Francisco, CA: Jossey-Bass.
- Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass
- Carr, J. W., & Padgett, T. C. (1992). Legal liability when designing or adopting student rating forms. In R. A. Arreola (Ed.), *Proceedings of the First Annual CEDA Conference on Practical Issues in Faculty Evaluation* (pp. 68-76). Memphis, TN: Center for Educational Development and Assessment.
- Centra, J. A. (1993). *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness*. San Francisco, CA: Jossey-Bass.
- Clark, S. J. (2003). *Use of online student ratings at institutions of higher education: Results of a web search*. Provo, UT: Brigham Young University Faculty Center.
- Franklin, J., & Theall, M. (1989, April). *Who reads ratings: Knowledge, attitudes, and practices of users of student ratings of instruction*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Services No. ED306241)
- Hmielewski, K., & Champagne, M. V. (2000, September/October). Plugging into course evaluation. *The Technology Source*. Retrieved July 25, 2006, from: http://technologysource.org/article/plugging_in_to_course_evaluation/
- Johnson, T. D. (2003). Online student ratings: Will students respond? In D. L. Sorenson & T. D. Johnson (Eds.), *New directions for teaching and learning: No. 96. Online student ratings of instruction* (pp. 49-60). San Francisco, CA: Jossey-Bass.
- Kaplin, W. A. (1978). *The law of higher education: A comprehensive guide to legal implications of administrative decision making*. San Francisco, CA: Jossey-Bass.
- Pascarella, E. T. (2001). Cognitive growth in college: Surprising and reassuring findings from The National Study of Student Learning. *Change*, 33(6), 21-27.
- Sorenson, D.L., & Johnson, T.D. (Eds.). (2003). *New directions for teaching and learning: No. 96. Online student ratings of instruction*. San Francisco, CA: Jossey-Bass.
- Theall, M. (2000, November/December) Electronic course evaluation is not necessarily the solution. *Technology Source*. Retrieved July 25, 2006, from: http://technologysource.org/article/electronic_course_evaluation_is_not_necessarily_the_solution/