Effective Evaluation of Teaching

A Guide for Faculty and Administrators

Edited by

MARY E. KITE

Table of Contents

Cover design by Haley Armstrong

Chapter Abstracts

Conducting Research on Student Evaluations of Teaching
William E. Addison & Jeffrey R. Stowell
The long and productive history of research on student evaluations of teaching (SETs) can be traced to the early 1920s. Following a summary of this history, we examine the methodologies and findings in four broad areas of research: reliability studies; validity studies; factor analyses; and investigations involving course, instructor, and student variables that have been examined for their possible influence on SETs. Additionally, we discuss methodological concerns and ethical issues associated with research in this area, and briefly describe several directions for future research.

Choosing an Instrument for Student Evaluation of Instruction
Jared W. Keeley
Student evaluations of instruction (SEIs) have become ubiquitous in the college classroom. The purpose of this chapter is to aid individuals in selecting an SEI to meet their particular evaluative goals. To do so, the chapter will review various considerations regarding the reliability, validity, and factor structure of SEIs and provide examples of publically available and for-pay instruments.

Formative Teaching Evaluations: Is Student Input Useful?
Janie H. Wilson and Rebecca G. Ryan
Student evaluations of teaching offer valuable information to teachers who want to improve their teaching. Specifically, formative student evaluations collected prior to the end of a course allow teachers to adjust teaching practices and potentially enhance learning. In this chapter, we discuss several characteristics of teaching evaluations, including content, timing, format, and ways to utilize evaluations effectively.

Using Student Feedback as One Measure of Faculty Teaching Effectiveness
Maureen A. McCarthy
Determining how to use Student Evaluations of Teaching (SETs) as a measure of teaching effectiveness has been a challenge for faculty and administrators alike. Quantitative measures, interpreted in isolation, provide succinct data that have the greatest potential for misinterpretation. In this chapter I provide recommendations for how to use both quantitative and qualitative feedback to improve instruction and to evaluate faculty effectiveness.

Bias in Student Evaluations
Susan A. Basow and Julie L. Martin
In this chapter, potential biasing factors in student evaluations of professors are examined. Because white male professors are the norm, faculty members who are female or who are from other racial/ethnic groups appear to be held to a higher or double standard of performance, Professor attractiveness ratings and age also affect student ratings, as do such course variables as expected grade. For example, higher expected grades are positively correlated with evaluations, even more so than actual grades. These findings should make us cautious in using student ratings as an unbiased measure of teaching effectiveness.

On-line Measures of Student Evaluation of Instruction
Cheryll M. Adams
In recent years, more institutions of higher education (IHE) have moved from paper and pencil surveys to online evaluations of instruction (Avery, Bryant, Mathios, Kang, & Bell, 2006). This practice has not eliminated the controversies such as whether students can effectively evaluate an instructor's teaching, but instead has brought new ones to the forefront. The advantages of using online evaluations include cost-effectiveness, more time for responding, and faster feedback to faculty. The trade off, in general, is a lower response rate for the evaluations. This chapter addresses research about online measures of instruction. The pros and con of using online measures of instruction instead of traditional paper and pencil measures are reviewed and recommendations are offered for using online measures of instruction effectively.

What's the Story on Online Evaluations of Teaching?
Michelle Drouin
As online teaching has gained popularity in the last decade, evaluations designed specifically for online teaching have begun to emerge. In this chapter, I give an overview of some of the most popular self, peer, and student evaluations of online teaching. I also discuss the different approaches to teaching outlined by Anderson & Dron (2011) (i.e., cognitive behavioral, social-constructivist, and connectivist) and give recommendations for rubrics that align with these different pedagogical approaches.

Using Course Portfolios to Assess and Improve Teaching
Paul Schafer, Elizabeth Yost Hammer, Jason Berntsen
This article describes how course portfolios can be used for both formative and summative assessments of teaching, and explains the difference between "teaching" and "course" portfolios. The article then details the successful Course Portfolio Working Group program at Xavier University of Louisiana, emphasizing the effectiveness of the program for the assessment and improvement of teaching. Practical advice is provided to assist individuals and institutions in the development of similar programs.

Peer Review of Teaching
Emad A. Ismail, William Buskist, and James E. Groccia
This chapter describes a formative model of peer review in which faculty observe other faculty teach in order to provide constructive feedback on their teaching. We outline a five- step process for peer review that includes a preclassroom visitation meeting, classroom observation of teaching, solicitation of student feedback, preparation of a written report, and a postclassroom visitation meeting with the teacher to provide formative feedback. We also address several questions and concerns that faculty often raise about peer review of college and university teaching.

Conducting Research on Student Evaluations of Teaching

William E. Addison & Jeffrey R. Stowell
Eastern Illinois University

Introduction

Research on student evaluations of teaching (SETs) has a long and fruitful history, producing several thousand publications (Feldman, 1997; Marsh, 2007). Some of the earliest work in the area was done by psychologist Max Freyd, who suggested that his graphic rating scale could be used to measure characteristics of the teacher that he accepted as "fundamental to the acquisition of a successful teaching technique" (1923, p. 434). According to Freyd, these characteristics include such attributes as alertness, sense of humor, tact, patience, acceptance of criticism, and interestingly, neatness in dress. Whether or not Freyd himself possessed these qualities is unclear; however, we do know that his teaching career at the University of Pennsylvania lasted just one year, after which he joined John Watson at the J. Walter Thompson Company in conducting research on marketing and advertising (Vinchur & Koppes, 2007).

Another early figure in research on SETs is Hermann Remmers, a professor of education and psychology at Purdue University. He and his colleague George Brandenburg conducted several studies using an instrument of their own design, the *Purdue Rating Scale for Instructors* (Brandenburg & Remmers, 1927; Remmers, 1928, 1930; Remmers & Brandenburg, 1927). Years later, Remmers collaborated on one of the earliest factor analyses of student evaluations (Smalzried & Remmers, 1943). For their study, Smalzried and Remmers used a version of the Purdue scale with just 10 items, or "traits." The 10 traits included such obvious qualities as "presentation of the subject matter" and "fairness in grading," as well as some less obvious ones, such as "personal appearance" and "personal peculiarities." Their analysis yielded two factors, which they called "professional maturity" and "empathy."

Prominent behaviorist Edwin Guthrie (1927) also conducted an early study of SETs in which he focused on differences in ratings given at the beginning of the academic year versus ratings given at the end of the year. In addition, he examined differences between student ratings of faculty and student *rankings* of faculty, concluding that ratings are more reliable than rankings. In 1949, when he was serving as Dean of the Graduate School at the University of Washington, he conducted a survey of faculty members that yielded, among other results, the finding that 18% of the faculty indicated that it is impossible or very difficult to measure teaching effectiveness.

Another important figure in the study of SETs is Wilbert McKeachie, professor emeritus at the University of Michigan and author of *Teaching Tips: Strategies, Research, and Theory for College and University Teachers*, first published in 1951 and currently in its 13th edition (Svinicki & McKeachie, 2010). Although McKeachie has published numerous articles on SETs over his lengthy career, his 1969 article in the *American Association of University Professors Bulletin* is particularly noteworthy for including the suggestion that when considered for the purpose of personnel evaluation, data collected from SETs should always be used in conjunction with other information about the teacher and the course.

Of all the researchers who have studied SETs, no individual has made a more significant contribution to the literature than Herbert Marsh, currently a professor at the University of Oxford. From the mid 1970s through the beginning of the 21st century, Marsh conducted dozens of studies on SETs, focusing on the instrument he designed, the SEEQ – Students' Evaluations of Educational Quality (e.g., Marsh, 1977,

1980, 1984, 1987; Marsh & Hocevar, 1991; Marsh & Roche, 1997, 2000). Among his most important findings regarding the SEEQ is that the instrument is reliable, relatively valid in comparison to other measures of teaching effectiveness, and generally unaffected by such potential biases as grading leniency and class size (see Keeley, this volume, for a summary of the psychometric properties of this instrument).

A relatively recent landmark in the study of SETs is the publication of a special issue of the *American Psychologist* in November, 1997 devoted to studies of SETs, including articles written by proponents as well as critics of student evaluations. Not surprisingly, both Marsh and McKeachie contributed to the issue and prominent social psychologist Anthony Greenwald, generally considered a critic of SETs, served as the action editor. Greenwald's affiliation with the University of Washington is worth noting in that it is the same institution at which Guthrie conducted his early research on SETs 5 decades earlier.

In his article, McKeachie (1997) suggested that the problems with SETs are generally not a function of the instruments themselves, but rather they are due mainly to how the ratings are used, especially the lack of sophistication among members of personnel committees who use student ratings to make promotion and tenure decisions. Consequently, he suggested that greater attention should be directed toward methods ensuring valid use of the ratings.

In the article he co-authored with his colleague Gerald Gillmore (Greenwald & Gillmore, 1997), Greenwald suggested that instructors' pursuit of positive ratings may lead to lenient grading, which in turn can diminish the academic content of courses. And like McKeachie, he cautioned against the misuse of SETs by personnel committees (e.g., by relying on means calculated across disparate dimensions of teaching).

Recent Trends

Since the late 1990s, the internet has played an increasingly important role in the use and study of SETs. The number of students enrolled in online courses is growing, with 29% of all college students taking at least one online course (Allen & Seaman, 2010). One of the challenges of teaching online classes is how to effectively evaluate these courses, especially because most SETs were developed for face-to-face classes (see Drouin, this volume). A related issue concerns the use of online student evaluations for all classes, including those that are taught face-to-face (see Adams, this volume, for a review).

Areas of Study

In this section of the chapter, we summarize the main findings in four broad areas of research on SETs: Reliability, validity, factor analyses, and variables that can influence student ratings. These areas of research are certainly not exhaustive, but studies that fall into one or more of these categories constitute the bulk of the research on SETs. We direct the reader to other chapters in this work, particularly those by Basow and Martin and by Keeley, to supplement the overview presented here.

Reliability studies

A large number of studies on SETs focus on the assessment of the reliability of the instruments. In these studies, researchers typically examine one of three types of reliability for a particular instrument: interrater reliability, stability, or generalizability (Hobson & Talbot, 2001). Not surprisingly, interrater reliability figures vary depending on the scale being used, but they also vary with the size of the class (the larger the class, the greater the reliability). For example, Cashin (1988) found reliability figures of .60 for a class of 5 students, .69 for a class of 10 students, .81 for a class of 20 students, and .95 for a

class size of 50. As Hobson and Talbot suggested, these findings indicate that a class size of at least 15 may be necessary for the meaningful use of SETs.

Researchers have also examined the stability of SETs over time. For example, Marsh (1984) used a cross-sectional approach to investigate the level of agreement between retrospective ratings from former students and those from current students and found substantial agreement between the ratings. Marsh and others have used longitudinal studies to examine the relationship between current students' evaluations and evaluations by the same students at least 1 year later (e.g., Overall & Marsh, 1980). Again, the findings indicate that these ratings are reasonably similar across time. Even over the course of 13 years, SETs of the same instructors tend to be quite stable (Marsh & Hocevar, 1991).

Studies of generalizability of SETs are typically designed to assess the extent to which student ratings reflect the effectiveness of a particular instructor rather than specific aspects of a course (e.g., subject matter, level, required vs. elective). The results from a number of studies in this area suggest that SETs tend to be highly generalizable across courses and students (e.g., Barnes & Barnes, 1993; Cashin, 1988).

Validity studies

Research on the validity of SETs generally falls into one of two categories: studies that examine the relationship between student ratings and measures of student learning, and those that compare SETs with other assessments of teaching effectiveness (e.g., peer ratings, self-ratings, expert ratings). Numerous researchers have examined the correlation between SETs and student learning as measured by grades, an area of research that has generated a considerable amount of controversy over the years. According to Aleamoni (1999), these studies have produced decidedly mixed results. In general, the correlations reported are either positive and weak, or virtually zero.Studies comparing SETs with other indices of instructor competence have generally yielded more consistent results. For example, based on an extensive review of the literature, Aleamoni and Hexner (1980) concluded that correlations between SETs and such measures as peer ratings, expert ratings, and alumni ratings tend to be moderate to high. But of course when it comes to research on SETs, there are always conflicting views. In this case, both Feldman (1988) and Marsh (1984) concluded that the correlation between SETs and peer ratings is relatively low.

Factor analyses

A number of studies of SETs have employed factor analysis in attempts to distill the various elements of effective teaching into a relatively small number of dimensions. As mentioned earlier, Smalzried and Remmers (1943) conducted one of the earliest studies of this kind, concluding that their 10-item scale could be condensed into 2 factors: “professional maturity” and “empathy.” According to Smalzried and Remmers, professional maturity includes items related to “the tools of the trade” (e.g., presentation of the subject matter), whereas empathy can be viewed as “the ability and willingness to wear each student’s sensorial and emotional shoes” (p. 366), and includes such qualities as fairness in grading and sympathetic attitude toward students.

About 20 years later, Robert Isaacson and his colleagues, including Wilbert McKeachie, conducted a factor analysis on data collected from approximately 300 introductory psychology students using an instrument that included 145 items (Isaacson, et al., 1964). Their analysis yielded six factors: (a) Overload (e.g., assigning a large amount of work); (b) Skill (e.g., explaining material clearly, stimulating intellectual curiosity); (c) Structure (e.g., following the syllabus, planning daily activities); (d) Feedback (e.g., providing comments, pro and con, on students’ work); (e) Group Interaction (e.g., encouraging

student participation); and (f) Rapport (e.g., listening attentively to students, providing reasons for criticism).

Another factor analysis comes from Peter Frey's (1978) study conducted at Northwestern University. Frey collected data from more than 26,000 student rating cards mailed to undergraduate students. Each card included seven statements on which students indicated their level of agreement. Examples of these statements are: "Class discussion was welcome in this course," and "The student was able to get personal help in this course." Interestingly enough, Frey's analysis revealed two factors that he called "pedagogical skill" and "rapport," which are virtually identical to the professional maturity and empathy factors identified by Smalzried and Remmers (1943).

In a study employing multiple factor analyses, Marsh (1991) examined data from more than 2000 evaluations, using his 35-item SEEQ. The results indicated that SEEQ responses could not be distilled into just a few factors, which is not surprising given that previous studies had consistently supported a nine-factor structure for the SEEQ (e.g., see Marsh, 1984; Marsh & Hocevar, 1984). The nine factors are: (a) Breadth of Coverage (e.g., discussed current developments); (b) Organization/Clarity (e.g., objectives stated and pursued); (c) Learning/Value (e.g., course is challenging, stimulating); (d) Examinations/Grading (e.g., exams were fair); (e) Enthusiasm (e.g., dynamic and energetic); (f) Rapport (e.g., interested in individual students); (g) Group Interaction (e.g., encouraged class discussion); (h) Assignments/Readings (e.g., readings were valuable); and (i) Workload/Difficulty (e.g., course workload was light/heavy). Although clearly less sophisticated than the factor analyses that Marsh used, an informal, "eyeball" analysis of the nine factors suggests that most of them can be included under the "skill" and "rapport" dimensions seen in earlier studies.

The overall results from factor analyses of SETs indicate that there are a limited number of factors associated with effective teaching and that these factors are generally related to one of two dimensions: how a course is designed and taught (the skills of teaching) and to what extent the instructor is able to establish a connection with students (rapport). Although there is some evidence that students tend to emphasize the rapport dimension over the skills dimension (Feldman, 1976), the independence of the dimensions suggests that the highest-rated instructors are those who, in the minds of students, effectively address both aspects.

Studies of Variables that May Influence SETs

A fertile area of research on SETs is the large number of variables that can potentially impact student ratings. In general, these studies can be classified into one of three categories, depending on whether the variable of interest is a characteristic of the course, the instructor, or the students. Among the more common course features to be examined by researchers for possible influence on SETs are class size, course content, course level, and whether students take the course as a requirement or an elective.

In terms of class size, the notion that students prefer smaller classes is generally supported in the literature (e.g., Aleamoni & Hexner, 1980; Mateo & Fernandez, 1996). However, other researchers have found no relationship between class size and student ratings (e.g., Shapiro, 1990), and some researchers have reported a curvilinear relationship between these variables in which both small and very large classes tend to produce more positive ratings (e.g., Marsh, Overall, & Kesler, 1979).

Researchers have also examined course content as a possible influence on SETs. For example, researchers have found that students taking classes related to their academic major tend to give relatively high ratings (e.g., Centra & Linn, 1976). Additionally, the amount of work combined with the

perceived difficulty of the class may influence student ratings (e.g., Babad, Avni-Babad, & Rosenthal, 2004; Marsh, 1980). Marsh (2001) found that students gave higher ratings to classes that they viewed as having mostly "valuable assignments" and few "unnecessary assignments." And not surprisingly, students tend to give lower ratings to instructors who teach unpopular required courses, such as statistics for students majoring in psychology (Gray & Bergmann, 2003).

Instructor characteristics such as rank/experience, research productivity, gender, and personality features may also influence SETs (see Basow & Martin, this volume). The findings from studies of the relationship between the instructor's rank or teaching experience and student ratings are mixed. Most early studies reported a positive correlation (e.g., Downie, 1952; Gage, 1961), whereas more recent studies have generally yielded nonsignificant correlations (e.g., Petchers & Chow, 1988). Feldman (1983) suggested that this relationship takes the form of an inverted U; that is, that ratings improve initially, peak at 6-8 years of teaching, and then gradually decline. It is noteworthy that Feldman's proposed peak of teaching effectiveness coincides roughly with the tenure decision at most institutions.

Studies of the relationship between research productivity and SETs have generally produced nonsignificant findings, although these results may depend on the instructor's academic discipline. For example, Centra (1983) reported consistent relationships between the number of publications and student ratings for teachers of social science courses, but not for teachers in other disciplines.

Similar to the results in other areas of research on SETs, the findings on the possible relationship between gender of the instructor and SETs are inconsistent, possibly due to potential interactions between gender and other factors such as academic discipline (see Basow & Martin, this volume). Although most studies report no gender differences (e.g., Addison & Tabb, 2004; Aleamoni, 1999), some studies have reported that male instructors are rated more positively than female instructors (e.g., Basow & Silbert, 1987; Summers, Anderson, Hines, Gelder, & Dean, 1996), and others have reported that female instructors are rated higher than male instructors (e.g., Tatro, 1995). In still other cases, researchers have found a type of interaction effect: female instructors tend to be rated higher on the rapport domain of teaching, and male instructors tend to be rated higher on the presentation and organization domains (e.g., Kierstead, D'Agostino, & Dill, 1988).

Such personality variables as enthusiasm and warmth may also influence SETs. For example, researchers have found that the more animated and enthusiastic the instructor is, the higher the student ratings (e.g., Basow & Distenfeld, 1985). Additionally, Best and Addison (2000) found that students' perceptions of the warmth of the instructor were positively correlated with student ratings in the rapport domain.

Of the research on the potential influence of student characteristics on SETs, by far the most attention has been devoted to studies of the relationship between students' ratings of instruction and their academic performance, as measured by grades. Although the general belief among faculty is that students tend to assign higher ratings to instructors of classes in which they expect or receive good grades, the preponderance of evidence does not support this view (Aleamoni, 1999). Based on an examination of dozens of studies, Aleamoni found a median correlation of approximately 0.14. As numerous researchers have suggested, we should not be surprised by positive correlations between SETs and measures of student learning, even if they are relatively weak correlations. In opposition to the oft-cited "leniency hypothesis," a perhaps more parsimonious explanation for this relationship is that students tend to perform better in classes taught by effective teachers.

Research Issues

Conducting research on student evaluations is not easy. Researchers face a number of challenges that include questions about the evaluation instrument itself (e.g., reliability and validity), the difficulty of selecting and assigning students or instructors to different experimental conditions, and ethical concerns that arise from methodologies that may alter students' perceptions and learning. Without consideration of these issues, results may be tenuous at best.

Methodological concerns

Perhaps the key methodological issues in any study of SETs are those that concern the reliability and validity of the instrument being used, which obviously depend on the items selected for inclusion on the instrument. The development of the Teacher Behavior Checklist (TBC) illustrates a "grass-roots" approach to developing an instrument for student evaluations, with a published history that serves as a prime example for overcoming concerns about the reliability and validity of an SET (see Keeley, this volume).

Following initial development of the TBC, researchers conducted several studies designed to assess its validity. For example, Buskist, Sikorski, Buckely, and Seville (2002) found considerable overlap between the list of qualities on the TBC and characteristics of master teachers identified through other methods, thus supporting the external validity of the TBC. Additionally, recent research suggests that the TBC is useful for discriminating between those who undergraduate students consider their best and worst teachers (Keeley, Furr, & Buskist, 2010). Finally, a factor analysis of the TBC indicated that the 28 qualities could be distilled into two factors: professional competency/communication skills, and caring/support (Keeley, Smith, & Buskist, 2006), results that are again remarkably similar to those obtained from other factor analyses of SETs.

Another area of concern related to conducting research on SETs is the extent to which extraneous factors can influence students' ratings. In addition to the trait variables we discussed earlier (course, instructor, and student characteristics) that can influence SETs, research in this area is susceptible to momentary or situational factors. For example, if the time of day can influence students' performance on a test (Hartley & Nicholls, 2008), it is possible that the time of day could also influence SETs. In fact, Youmans and Jee (2007) deliberately controlled the time of day and other potential situational confounds such as the day of the week and time to complete the evaluations while manipulating whether or not students were offered chocolate prior to completing the evaluations. Their purpose in doing so was to eliminate potential external influences on student evaluations except for their variable of interest (i.e., the effects of chocolate). Interestingly, chocolate had a small to moderate enhancing effect on student ratings. More importantly, Youmans and Lee recommended that the reliability and validity of SETs could be improved by standardizing the method of administration and being cautious in interpreting the results from any one sample of evaluations.

The potential impact of extraneous variables is particularly problematic for studies of online evaluations, given that they can be completed by students anytime and anywhere. In fact, this is one of the common concerns that instructors have about online evaluations. As Stowell, Addison, and Smith (2011) suggested, students could be completing evaluations under different situational contexts including late at night, under the influence of alcohol, in collaboration with their peers, or after enjoying a pleasant spring day outside. Some of these external influences could also influence classroom-based evaluations, but the online setting potentially increases the amount of error variance when trying to assess the effects of a treatment on student evaluations (see Adams, this volume).

As with other research on teaching and learning, two other methodological challenges arise in studies of SETs: the difficulty of obtaining a random sample and being able to randomly assign participants to control and experimental groups. Random sampling, in which each person in the population has an equal chance of participating, is important for being able to generalize the results from the sample to the population of interest (e.g., all students). Random assignment, the use of a method of chance to determine which participants are in the control and experimental groups, reduces bias in experimental designs by ensuring that the groups are relatively similar in background characteristics (e.g., sex, race, motivation, intelligence, etc.) prior to any experimental manipulation.

Suppose an instructor were interested in knowing if activities on the first day of class influence students' impressions of the instructor and the class that persist long enough to influence instructor evaluations at the end of the course. Ideally, a researcher would gather a random sample of participants from the student population that includes students from various backgrounds, randomly assign them to the experimental and control conditions, and measure outcomes on student evaluations. Unfortunately, random selection from the population is virtually impossible, as instructors typically do not have control over who enrolls in their courses. However, instructors may be able to randomly assign students to different conditions using a quasi-experimental design in which separate sections of the same course (ideally taught by the same instructor) are randomly assigned to the experimental and control groups. Hermann, Foster, and Hardin (2010) used this approach to find that introductory psychology students (taught by different instructors) who experienced a first-day reciprocal interview gave higher ratings of course satisfaction at the end of the semester.

Stowell et al. (2011) used a similar multi-section sample to examine differences in SETs conducted online in a face-to-face course compared to SETs completed in the classroom. They searched class schedules to find instructors who were teaching more than one section of the same course and then assigned each section to the experimental or control condition, counterbalancing the order of the sections to avoid time-of-day or practice effects from one section to the next. Interestingly, the format of the SETs did not affect mean ratings or type of comments provided by students.

Ethical issues

As with other areas of research (e.g., health and medicine), sometimes the preferred research design (i.e., use of random sampling and random assignment) creates ethical concerns. For example, research on the relationship between grades and instructor ratings suggests that instructors might be able to improve their evaluations by simply making the course easier and assigning higher grades (Greenwald & Gillmore, 1997). Obviously, it would be difficult to test this premise experimentally due to the ethical issues involved in awarding students grades different from what they earned. However, Worthington and Wong (1979) used a creative approach to examine this hypothesis while minimizing the ethical issues. After 2 weeks of lectures on a psychology topic, they had students complete a multiple choice quiz on which they received feedback of "good," "satisfactory," or "poor," regardless of their actual quiz performance. Students then completed instructor evaluations for the course, followed by a debriefing session. Worthington and Wong found that actual quiz grades were not strongly related to the instructor ratings, but the instructor ratings were highest among students who performed poorly and received an artificially inflated grade. As an alternative to using deception, Howard and Maxwell (1982) used a longitudinal design to study grades and satisfaction with the course, finding a weak relationship between the variables.

Another ethical issue concerns how changes in instructor ratings could be an inadvertent casualty of conducting research in the scholarship of teaching and learning. Instructors who are willing to take risks

by introducing new methods into their teaching are likely to experience a change in instructor ratings as a result of a change in their pedagogy, which could potentially affect tenure, promotion, and salary (see McCarthy, this volume, for discussion of SETs in faculty evaluations). Secondly, if an instructor's research design includes a no-treatment control group, it may be considered unethical to knowingly withhold a pedagogical technique that could improve student learning. One solution to this problem is to implement a within subjects pre-post design, but include alternative forms of the assessments assigned to each group at each time point (e.g., Bartsch, Engelhardt Bittner, & Moreno, 2008). Applying this method in the context of SET research would entail finding two similar assessments of teaching effectiveness (or splitting one instrument in two halves), randomly assigning students to take one of the pretests, and then give each student the alternate form after all students experience the treatment (i.e., teaching method). A significant main effect of time (pre/post) supports the effectiveness of the treatment (Bartsch et al., 2008).

Directions for Future Research

One area of study that has seen an increased amount of attention in recent years is the perception of the evaluation process, by faculty as well as students. For example, Balam and Shannon (2010) used Aleamoni's "myths" about student ratings (see Aleamoni, 1987, 1999) as the basis for a study of the differences between students' and faculty members' perceptions of SETs. In general, they found that students were more likely than faculty to believe the myths, although faculty held stronger beliefs in the myth that student ratings are unreliable and invalid. In another recent study, Heine and Maddox (2009) found gender differences in students' views of the evaluation process. Among their more interesting findings is that male students were more likely than female students to believe that instructors adjusted their in-class behavior at the end of the semester in order to attain higher ratings.

One possible direction for future research is to examine the relative influence that rapport and teaching skill have on SETs. As we discussed earlier, a number of researchers have suggested that the dimensions of effective teaching can be reduced to two broad factors: what Smalzried and Remmers (1943) called "professional maturity" and "empathy," and what Frey (1978) called "pedagogical skill" and "rapport." Similarly, Lowman (1995) suggested that these two factors are related to intellectual stimulation (e.g., clarity of presentation, enthusiasm for content) and interpersonal rapport, which is based on the instructor's display of respect and care for students. Research involving these two dimensions of teaching has typically been conducted either as components of more general studies on the multidimensionality of effective teaching (e.g., Buskist, Sikorski, Buckley, & Saville, 2002; Frey, 1978), or as studies of the specific role that rapport plays in the teaching and learning process (e.g., Benson, Cohen, & Buskist, 2005; Wilson, Ryan, & Pugh, 2010). For example, Wilson and her colleagues found that a measure of rapport between teachers and students significantly predicted students' motivation, their perceptions of learning, and their self-reported grades.

Advances in technology (e.g., online evaluations, the use of student response systems or "clickers") also provide opportunities for original research. As new technologies are developed, they are likely to play an increasingly important role in teaching and learning. Consequently, there will be a need for research on how these yet-unknown technologies impact the assessment of teaching effectiveness.

In sum, nearly a century's worth of research on student evaluations has demonstrated that, depending on the instrument, SETs are multidimensional, generally reliable, relatively valid when compared to other measures of teaching effectiveness, and primarily a function of the teacher rather than the course. However, given the number of different evaluation instruments in use, the inconsistency of findings in many areas, the multitude of variables that seem to influence SETs, and the importance of student

ratings in promotion and tenure decisions, it is likely that SETs will continue to be a popular focus of study for the foreseeable future.

References

Addison, W. E., & Tabb, S. L. (July, 2004). *Do instructions to students and gender of instructor influence evaluations of teaching*? Poster session presented at the meeting of the American Psychological Association, Honolulu.

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

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153-166.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science*, *9*, 67-84.

Allen, I. E., & Seaman, J. (2010). *Class differences: Online education in the United States, 2010.* Retrieved from http://sloanconsortium.org/publications/survey/class_differences

Babad, E., Avni-Babad, D., & Rosenthal, R. (2004). Prediction of students' evaluations from brief instances of professors' nonverbal behavior in defined instructional situations. *Social Psychology of Education*, *7*, 3-33. doi: 10.1023/B:SPOE.0000010672.97522.c5

Balam, E. M., & Shannon, D. M. (2010). Student ratings of college teaching: A comparison of faculty and their students. *Assessment & Evaluation in Higher Education*, *35*(2), 209-221. doi: 10.1080/02602930902795901

Barnes, L. L. B., & Barnes, M. W. (1993). Academic discipline and generalizability of student evaluations of instruction. *Research in Higher Education*, 34, 135-149.

Bartsch, R. A., Engelhardt Bittner, W. M., & Moreno, J. E. (2008). A design to improve internal validity of assessments of teaching demonstrations. *Teaching of Psychology*, *35*(4), 357-359. doi:10.1080/00986280802373809

Basow, S. A., & Distenfeld, M. S. (1985). Teacher expressiveness: More important for male teachers than female teachers? *Journal of Educational Psychology*, *77*, 45-52. doi: 10.1037/0022-0663.77.1.45

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology*, *79*(3), 308-314. doi: 10.1037/0022-0663.79.3.308

Benson, T. A., Cohen, A. L., & Buskist, W. (2005). Rapport: Its relation to student attitudes and behaviors toward teachers and classes. *Teaching of Psychology*, *32*(4), 237-239. doi: 10.1207/s15328023top3204_8

Best, J. B., & Addison, W. E. (2000). A preliminary study of perceived warmth of professor and student evaluations, *Teaching of Psychology*, *27*, 60-62.

Brandenburg, G. C., & Remmers. H. H. (1927). A rating scale for instructors, *Educational Administration and Supervision, 13*, 399-406.

Buskist, W., Sikorski, J., Buckley, T., & Saville, B. (2002). Elements of master teaching. In S. F. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 27-39).

Cashin, W. E. (1988). Student ratings of teaching: A summary of the research. (IDEA paper no. 20). Manhattan: Kansas State University, Center for Faculty Evaluation and Development. Retrieved from http://www.theideacenter.org/sites/default/files/Idea_Paper_20.pdf

Centra, J. A. (1983). Research productivity and teaching effectiveness. *Research in Higher Education*, *18*(4), 379-389.

Centra, J. A., & Linn, R. L. (1976). Student points of view in ratings of college instruction. *Educational and Psychological Measurement*, *36*, 693-703.

Downie, N. M. (1952). Student evaluations of faculty. *Journal of Higher Education*, *23*, 495-496, 503.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education*, *5*, 243-288.

Feldman, K. A. (1983). The seniority and instructional experience of college teachers as related to the evaluations they receive from their students. *Research in Higher Education*, *5*(3), 243-288. doi: 10.1007/BF00991967

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education*, *28*(4), 291-344.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds), *Effective teaching in higher education: Research and practice* (pp. 93-143). New York: Agathon Press.

Frey, P. W. (1978). A two-dimensional analysis of student ratings of instruction. *Research in Higher Education*, *9*, 69-91. doi: 10.1007/BF00979187

Freyd, M. (1923). A graphic rating scale for teachers. *Journal of Educational Research*, *8*(5), 433-439.

Gage, N. L. (1961). The appraisal of college teaching. *Journal of Higher Education*, 32, 17-22.

Gray, M., & Bergmann, B. R. (2003). Student teaching evaluations: Inaccurate, demeaning, misused. *Academe*, *89*, 44-46.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*(11), 1209-1217. doi: 10.1037/0003-066x.52.11.1209

Guthrie, E. R. (1927). Measuring student opinion of teachers. *School and Society*, *25*, 175-176.

Guthrie, E. R. (1949). The evaluations of teaching. *The Educational Record*, *30*, 109-115.

Hartley, J., & Nicholls, L. (2008). Time of day, exam performance and new technology. *British Journal of Educational Technology*, *39*(3), 555-558. doi: 10.1111/j.1467-8535.2007.00768.x

Heine, P., & Maddox, N. (2009). Student perceptions of the faculty course evaluation process: An exploratory study of gender and class differences. *Research in Higher Education Journal*, *3*, 1-10. Retrieved from http://aabri.com/manuscripts/09192.pdf

Hermann, A. D., Foster, D. A., & Hardin, E. E. (2010). Does the first week of class matter? A quasi-experimental investigation of student satisfaction. *Teaching of Psychology*, *37*(2), 79 - 84. doi: 10.1080/00986281003609314

Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching*, *49*, 26-31.

Howard, G. S., & Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education*, *16*(2), 175-188. doi: 10.1007/bf00973508

Isaacson, R. L., McKeachie, W. J., Milholland, J. E., Lin, Y. G., Hofeller, M., Baerwaldt, J. W., & Zinn, K. L. (1964). Dimensions of student evaluations of teaching. *Journal of Educational Psychology*, *55*, 344-351. doi: 10.1037/h0042551

Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the teacher behavior checklist. *Teaching of Psychology*, *37*(1), 16 - 20. doi:10.1080/00986280903426282

Keeley, J., Smith, D., & Buskist, W. (2006). The teacher behaviors checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology*, *33*(2), 84-91. doi: 10.1207/s15328023top3302_1

Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology*, *80*(3), 342-344. doi: 10.1037/0022-0663.80.3.342

Lowman, J. (1995). *Mastering the techniques of teaching* (2nd ed.). San Francisco: Jossey-Bass.

Marsh, H. W. (1977). The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. *American Educational Research Journal*, *14*, 441-447. doi: 10.2307/1162341

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. *American Educational Research Journal*, *17*, 219-237. doi: 10.2307/1162484

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707-754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*, 253-388.

Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, *83*, 285-296. doi: 10.1037/0022-0663.83.2.285

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal*, *38*, 183-212. doi: 10.3102/00028312038001183

Marsh, H. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383): Springer Netherlands.

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal*, *21*, 341-366. doi: 10.2307/1162448

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education*, *7*(4), 303-314. doi: 10.1016/0742-051x(91)90001-6

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Class size, student evaluations, and instructional effectiveness. *American Educational Research Journal*, *16*, 57-69.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*, 1187-1197. doi: 10.1037/0003-066X.52.11.1187

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myths, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92*(1), 202-228. doi: 10.1037/0022-0663.92.1.202

Mateo, M. A., & Fernandez, J. (1996). Incidence of class size on the evaluation of university teaching quality. *Educational and Psychological Measurement*, *56*(5), 771-778.

McKeachie, W. J. (1969). Student ratings of faculty. *AAUP Bulletin*, *55*, 439-444.

McKeachie, Wilbert J. (1997). Student Ratings: The Validity of Use. *American Psychologist*, 52(11): 1218-1225. doi: 10.1037/0003-066X.52.11.1218

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology*, *72*, 321-325. doi: 10.1037/0022-0663.72.3.321

Petchers, M. K., & Chow, J. C. (1988). Sources of variation in students' evaluations of instruction in a graduate social work program. *Journal of Social Work Education*, *24*(1), 35-42.

Remmers, H. H. (1928). The relationship between students' marks and students' attitudes toward instructors. *School and Society*, *28*, 759-760.

Remmers, H. H . (1930). To what extent do grades influence student ratings of instructors? *Journal of Educational Research*, *21*, 314-316.
Remmers, H. H., & Brandenburg, G. C. (1927). Experimental data on the Purdue Rating Scale for Instructors. *Educational Administration and Supervision*, *13*, 519-527.
Schaeffer, G., Epting, K., Zinn, T., & Buskist, W. (2003). Student and faculty perceptions of effective teaching: A successful replication. *Teaching of Psychology*, *30*(2), 133-136.
Shapiro, E. G. (1990). Effect of instructor and class characteristics on students' class evaluations. *Research in Higher Education*, *31*, 135-148.
Smalzried, N. T., & Remmers, H. H. (1943). A factor analysis of the Purdue Rating Scale for Instructors. *Journal of Educational Psychology*, *34*, 363-367. doi: 10.1037/h0060532
Stowell, J. R., Addison, W. E., & Smith, J. L. (2011). Comparison of online and classroom-based student evaluations of instruction. *Assessment & Evaluation in Higher Education*, 1-9. doi: 10.1080/02602938.2010.545869
Summers, M. A., Anderson, J. L., Hines, A. R., Gelder, B. C., & Dean, R. S. (1996). The camera adds more than pounds: Gender differences in course satisfaction for campus and distance learning students. Journal of Research and Development in Education, 29(4), 212-219.
Svinicki, M., & McKeachie, W. J. (Eds.). (2010). McKeachie's teaching tips : strategies, research, and theory for college and university teachers. Belmont, CA: Wadsworth, Cengage Learning.
Tatro, C. N. (1995). Gender effects on student evaluations of faculty. *Journal of Research and Development in Education*, 28(3), 169-173.
Vinchur, A. J., & Koppes, L. L. (2007). Early contributors to the science and practice of industrial psychology. In L. L. Koppes (Ed.), *Historical perspectives in industrial and organizational psychology* (pp. 37-58): Erlbaum, Mahwah, NJ.
Wilson, J. H., Ryan, R. G., & Pugh, J. L. (2010). Professor-student rapport scale predicts student outcomes. *Teaching of Psychology*, *37*(4), 246-251. doi: 10.1080/00986283.2010.510976
Worthington, A. G., & Wong, P. T. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology*, *71*(6), 764-775. doi: 10.1037/0022-0663.71.6.764
Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology*, *34*(4), 245 - 247. doi: 10.1080/00986280701700318

Contact Information

William E. Addison
Psychology Department
Eastern Illinois University
600 Lincoln Avenue
Charleston, IL 61920
Phone: 217-581-6417
Email: weaddison@eiu.edu

Jeffrey R. Stowell
Psychology Department
Eastern Illinois University
600 Lincoln Avenue
Charleston, IL 61920
Phone: 217-581-2279
jrstowell@eiu.edu

Choosing an Instrument for Student Evaluation of Instruction

Jared W. Keeley
Mississippi State University

Student evaluations of instruction (SEIs) have become ubiquitous in the college classroom. At many institutions, they have become synonymous with the evaluation of teaching, although, as this volume demonstrates, there is certainly more to the process of assessing teaching quality. Nonetheless, given their prominence, it is important to consider a variety of issues when selecting an SEI for use in your course, department, or institution. This chapter will outline a few of those critical concerns while providing examples of some currently available public-domain and fee-based instruments.

General Issues to Consider

Perhaps the first issue to consider when selecting an SEI is how you are planning to use it. Within higher education, there are many purposes for gathering information from students about your teaching. Generally, these purposes are divided into two broad categories termed summative versus formative evaluations. Summative evaluations serve administrative purposes, such as providing evidence of teaching effectiveness for tenure and promotion decisions or providing a metric for merit raises. On the other hand, formative evaluations serve the purpose of improving one's teaching. As such, formative evaluations are much more personal in nature and might be more explicitly focused upon a question one has about a course, such as "How well did this new demonstration work?" This chapter will focus mainly upon summative purposes when selecting an SEI, as these purposes tend to be much more uniform across institutions and individuals than formative questions. Nonetheless, formative and summative purposes are not mutually exclusive, and ideally a chosen evaluation plan should address aspects of both. For a more in-depth analysis of formative evaluations, please refer to Wilson and Ryan (this volume); for a discussion of how teaching portfolios can be used for this purpose, see Schafer, Hammer, and Berntsen (this volume).

Once instructors have considered the purposes for your evaluation, there are a variety of concerns about the psychometric qualities of the instrument. Traditionally, the psychometric evaluation of self-report instruments has highlighted the factors of reliability and validity. Reliability addresses the consistency of the measurement whereas validity refers to the accuracy with which the measurement captures what it was intended to measure. These concepts raise special concerns relative to student evaluation of instruction, and so I will address them in some depth.

Reliability

First, regarding reliability, it can be quite difficult to establish the consistency over time of student ratings of their instructor because those ratings are designed to change. For example, the "testing effect" (Roediger & Karpicke, 2006), where one administration of an instrument, in and of itself, changes the result of a second administration, has a pronounced influence upon SEIs. Therefore, simply administering an evaluation at one point tends to increase student ratings at a second point (Keeley, Smith, & Buskist, 2006). Therefore, reliability must often be measured in relative terms (e.g., all people who rate a 4 the first time change a consistent amount) rather than in absolute terms (e.g., a rating of 4 stays a 4).Thus, traditional test-retest reliability can be expected to be quite low, even across short time frames (like 2 weeks, but see Marsh, 1982a for an exception).

Therefore, a more preferred metric of reliability for SEIs is internal consistency, or how consistent individuals are across items within the same administration. However, again this metric can be difficult to apply to SEIs because of the nature of ratings. There tend to be pronounced halo effects within SEIs (Feeley, 2002; Marsh, 1984). For example, if a student likes the professor, he or she tends to rate the professor highly on all items (or vice versa). The result of a halo effect is that internal consistency is artificially inflated. In the case where SEIs accurately reflect differences in teacher characteristics across items (i.e., the variance we want to measure), internal consistency values will drop. Therefore, perfect internal consistency values would be an undesirable characteristic; instead medium to high values are ideal ($\alpha \approx .60$-$.80$)

Despite these limitations to measuring reliability within SEIs, nonetheless it is possible to compare various SEIs in terms of their test-retest and internal consistency values because they all face the same measurement issues. Unlike some other measurement domains, it is not possible to recommend cutoffs (e.g., $r = .90$) for acceptable reliability because of the problems noted above. However, the relative reliabilities of similarly tested SEIs can be compared, such that it is possible to say that one instrument is more reliable than another and such considerations should play a role in one's selection of an SEI.

Validity

The validity of SEIs is similarly fraught with pitfalls. Traditional psychometrics break the concept of validity into several domains, such as content validity, face validity, predictive validity, construct validity, etc. However, theorists hold that all validity evidence can be reduced to construct validity (Cronbach, 1980), or the accuracy of the measurement relative to the hypothetical construct it is designed to emulate. In this case, the hypothetical construct might be "good teaching," which could have a very multifaceted definition (Chism, 2004; Elton, 1998; Kreber, 2002; Marsh, 1984). Instead, we might focus upon lower level constructs of good teaching like communication skills or fairness in grading. It is important to note that the validity of an instrument is always tied to its purpose. A particular SEI might be valid for one purpose but not for another, depending upon its construction. For example, if the developers of an SEI did not consider interpersonal factors like rapport or immediacy to be important in their definition of "good teaching," then those constructs would be poorly represented in their measure. If part of your purpose of evaluation was to judge the nature of the student-instructor relationship, that instrument would have poor validity for that purpose. Therefore, an important aspect of selecting an SEI is that its development matches an instructor's purpose. Specific examples are discussed below, but it is important to review the initial development of the measure to ensure that it will offer adequate coverage of its intended purpose.

An often criticized aspect of validity for SEIs has been their relationship to outcome measures like grades. Ideally, all aspects of "good teaching" as a construct should have some relationship to how students perform in a class. However, the relationship is not so clear-cut. There appear to be confounding effects with grades, such that instructors are rated higher when students receive higher grades. It could be that the relationship is explained by students expressing satisfaction with more lenient grading policies (i.e., "I like you because you gave me an A."). Alternatively, the higher grades could be a reflection of better teaching practices, which lead to more student knowledge and/or learning, and therefore higher student performance on tests and assignments. Regardless of the nature of the effect (which is hotly debated; Centra, 2003; Gump, 2007; Langbein, 2008; Smith, Cook, & Buskist, 2011), grades do have an effect upon SEI results. In either case, the issue with grades highlights the fact that students are assumed to have the capability to comment upon "good teaching" practices, and almost certainly students' definition of those practices vary. Refer to Basow and Martin, this volume, for a discussion of the potential for gender and other biases in student ratings. For these reasons, SEIs

should never be one's only strategy for gathering information about one's teaching (for complimentary strategies, see Shafer et al., this volume.).

Factor Structure

A final psychometric consideration that is not often addressed is the factor structure of the instrument in question. In traditional psychometrics, a factor is a latent (that is, not observable) variable that represents the communality between a set of items. For example, if an instrument has 10 different items that all assess some aspect of the relationship between the student and instructor, and those items all correlate amongst themselves moreso than with other items on the scale, there might be an underlying factor (e.g., rapport) that influences students' ratings across those items. The factor structure of an SEI might be explicitly developed (e.g., drafting a set of items intended to measure an underlying aspect of teaching and then refining that set of items until the measurement is "clean") or it could be emergent (e.g., items were not developed with any intention for capturing an underlying factor, but such a structure emerges upon later examination).

Interestingly, many SEIs have evidenced similar factor structures regardless of whether they were intentionally developed or not. Two general factors that represent the procedural skills of teaching (e.g., being an effective communicator, constructing fair assignments) and the interpersonal aspects of teaching (e.g., being approachable, warm, considerate) appear to emerge consistently in the literature (Addison, 2005; Barnes et al., 2008; Keeley et al., 2006; Lowman, 1995). Although a variety of subconstructs have been identified, all aspects of SEIs can usually be placed into one of those general categories. As such, when selecting an SEI, it might be useful to ensure that aspects of the procedural and interpersonal factors are equally represented.

Publicly Available Instruments

A wide variety of SEIs have been developed and most of these instruments are "home-grown," meaning that they are developed at a particular institution for use at that institution. Unfortunately, the psychometric properties of these instruments are often unknown. However, a few instruments have been published within the teaching literature, along with evidence of the instrument's reliability, validity, and factor structure. These instruments are publicly available for use and therefore a number of individuals and institutions have adopted them as part of their evaluation package. A review of three of these instruments follows. By no means are these three the only available instruments; however, they are good examples of the development principles outlined above.

Student Evaluation of Educational Quality (SEEQ)

The Student Evaluation of Educational Quality (SEEQ; Marsh, 1982a) is perhaps the most investigated SEI in the pedagogical literature, largely by virtue of being one of the oldest. It grew from an initiative at the University of California Los Angeles (UCLA) and the University of Southern California (USC) to develop a standard, empirically sound evaluation instrument. The SEEQ consists of 35 items chosen from a larger pool developed from reviews of the literature, other evaluation forms, and interviews with faculty and students (Marsh, 1982a; 1984). Items were selected if students rated them as important, faculty rated them as useful, they were internally reliable, and they loaded with other similar items in factor analysis. These items were then submitted to extensive evaluation over hundreds of instructors and courses and thousands of students (Marsh, 1982a; 1982b; Marsh & Hocevar, 1984). The initial development of the SEEQ identified nine factors: (a) learning/value, (b) enthusiasm, (c) organization, (d) group interaction, (e) individual rapport, (f) breadth of coverage, (g) examinations/grading, (h) assignments, and (i) workload/difficulty (Marsh, 1982a). This initial structure was replicated by independent exploratory factor analysis of instructors at USC (Marsh & Hocevar, 1984), an Australian

sample of instructors (Marsh, 1981), and instructors self-ratings (Marsh, 1982b). Although the series of replications suggests a stable structure, it should be noted that the data for the first and third replication partially overlapped with the data used to generate the initial structure, which could explain the similarity of the findings. Further, the investigators did not examine alternative possible factor structures, and thus it is possible that another structure (such as a one or two factor solution) could also be viable.

Nonetheless, the reliability and validity of the SEEQ have been impressive. Internal consistency coefficients (alpha) range between .87 and .98 for the different factors (Marsh, 1982a; 1982b). Ratings appear stable over time; a sample of students were asked to retrospectively rate courses several years later and these ratings correlated with end-of-term reports at .83, indicating solid test-retest reliability (Marsh, 1982a). However, to date estimations of test-retest reliability have not been completed for shorter time frames, such as from mid-to-end semester. There is some evidence that the factor structure does not change across course offerings (Marsh & Hocevar, 1984). The SEEQ has demonstrated convergent validity in that ratings of individual traits (factors on the SEEQ) by different sources (instructors and students) for the same course tend to have higher relationships than do cross-trait correlations either within or across sources (Marsh, 1982b). However, the SEEQ appears to be subject to substantial halo effects, whereby student ratings across the nine factors are correlated and substantial variance is due to the rating method (i.e., students provide systematically different ratings than instructors do of themselves; Marsh, 1982b, 1984).

Teacher Behavior Checklist (TBC).

The Teacher Behavior Checklist (TBC) came from Buskist and colleagues' (2002) investigation of the traits of "master teachers." In their study, students listed qualities of master teachers, resulting in a list of 47 characteristics. A separate group of undergraduate students then generated behaviors that corresponded to those characteristics in an effort to operationalize how those characteristics are observed by students in the classroom. The list of behaviors in many cases overlapped across characteristics so the list was reduced to 28 items. In a new sample, both students and teachers rank ordered the importance of these 28 qualities There was substantial overlap between the top choices of students and faculty, but students placed more importance upon interpersonal factors of teaching and faculty emphasized the "nuts-and-bolts" technical aspects of teaching. The results of Buskist et al. (2002) have been replicated in American, Canadian, and Japanese community colleges, public universities, and private schools (Epting, Zinn, Buskist, & Buskist, 2004; Keeley, Christopher, & Buskist, in press; Schaeffer, Epting, Zinn, & Buskist, 2003; Vulcano, 2007; Wann, 2001).

These 28 items were converted into an SEI with the addition of a 5-point rating scale indicating how frequently (always to never) the teacher engaged in those behaviors. An initial factor analysis suggested either a one or two factor solution. The two factor option corresponded to two scales termed "caring and supportive" items versus "professionalism and communication skills" items (Keeley et al., 2006). The single factor solution was assumed to correspond to a general "good teaching" construct. A later confirmatory factor analysis found that a hybrid factor model with two subscales and an overarching total scale provided the best fit to the data. The internal consistency (coefficient alpha) of the two subscales was .93 and .90, respectively, with a coefficient of .95 for the total scale. Test-retest estimates of reliability from the middle to end of the semester were .68 for the caring and supportive scale, .72 for the professionalism and communication skills scale, and .71 for the total scale, with the understanding that actual value of the ratings increased an average of half a point across the term (Keeley et al., 2006).

The TBC appears to differentiate between teachers of differing ability. In one study, TBC ratings corresponded to an identical pattern of differences on a standard university evaluation form for four professors (Keeley et al., 2006). In another set of studies, students rated the best professor they had ever had, the worst professor they had ever had, and the professor from whom they most recently had class. Utilizing generalizability analysis (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Furr & Bacharach, 2008), it is possible to determine if individual items or students have a differential effect upon TBC scores. Students responded consistently across the three ratings of teacher type, such that the quality of the teacher (best, worst, most recent) had the only major effect upon TBC score differences. In other words, the TBC appears to provide a relatively clean measurement of teacher quality (Keeley, Furr, & Buskist, 2010).

Barnes et al.'s Unnamed Measure

Another example of a well developed, psychometrically evaluated instrument is one developed by Barnes and colleagues (2008). They began by selecting a panel of eight expert judges who reviewed items from nine commonly used SEIs. When the highest quality items had been selected from that initial pool, two judges sorted the items into categories and six judges independently rated how well those items exemplified those categories. Some new items were generated so that each category had adequate coverage. They considered the items and categories to reflect two general domains: *teaching readiness* which incorporated knowledge base, teaching skill, and professionalism, and *teaching excellence*, which incorporated interpersonal aspects of teaching like rapport and enthusiasm.

The item pool was further refined through factor analysis, such that items that did not load on the intended two factors were eliminated. Additionally, items that did not produce a sufficient range of response on a 7-point rating scale were removed, as they would not contribute meaningfully to measurement variance. The final scale included 6 items of teaching readiness with a Cronbach's alpha of .88 and 8 items of teaching excellence with a Cronbach's alpha of .95 (Barnes et al., 2008). Confirmatory factor analysis indicated that the two factor structure provided an adequate fit to the data. They found that scores on each of the scales were significantly related to students' expected grades, such that higher evaluations were correlated with higher grades. In addition, the teaching readiness score seemed to result in higher ratings than the teaching excellence score across the board, which the authors took as evidence of the validity of the ratings, as achieving high ratings excellence should be more difficult for teachers (Barnes et al., 2008).

Private Evaluation Systems

In addition to published, publicly available instruments, a number of private systems also exist, either at universities or within companies. In contrast to the publically available instruments, which tend to be a single set of questions, the private systems can be much more complicated. Again, there are many such systems; the two reviewed below are intended to be good examples of some common options.

University of Washington

The Office of Educational Assessment at the University of Washington (2005) has developed an extensive evaluation system over the course of many years. The system consists of a set of evaluation forms rather than a single instrument. Forms are available for different course and section types, such as large lectures, small discussions, lab sections, distance learning courses, English as a second language, and skills acquisition. Each form has slightly different questions depending upon the nature of the course. Instructors are able to select which forms are most applicable to their course format. This system is more flexible than a single instrument, as some items on a general instrument might not apply in particular contexts.

The forms were developed through a purely content validity orientation. Items were developed through a series of interviews with faculty, administrators, and student focus groups about their courses and assessment needs. Drafts of the items were revised based upon feedback from both students and faculty. Otherwise, no validity data are available regarding the forms. Internal consistency coefficients range from .75 to .90 on individual items across the forms, with some forms indicating greater reliability than others (University of Washington, 2005). The reliability of items across courses seems to be adequate when aggregating at least seven courses and plateaus at around fifteen courses (Gillmore, 2000). Gillmore and Greenwald (1994) noted significant halo effects with students responding generally positively towards courses, even when trying to control for comparison to other courses.

The IDEA Center

The Individual Development and Educational Assessment (IDEA) was originally developed by Hoyt (1973). It currently is administered by an independent non-profit organization of the same name. A substantial revision process occurred in 1997 when new learning objectives and methods were added to current IDEA items. Within this system, an instructor first conducts a self-assessment of the course, identifying from a predetermined list particular objectives for a course (e.g., knowledge base, critical thinking skills, working with a group). These objectives are then included as measurements in the SEI to determine the degree to which students feel that they have reached those goals over the course of the term. The SEI also includes students' ratings of the methods used to reach those objectives. The methods items were selected empirically by determining, through stepwise regression, which items significantly and independently predicted outcome objectives and only those items were included in the revised IDEA system. The IDEA system also statistically controls for a number of extraneous variables: class size, student motivation, discipline-related difficulty, and student effort. For example, students tend to rate the instruction more highly in smaller courses. The IDEA system statistically controls for these effects upon ratings (Hoyt, Chen, Pallett, & Gross, n.d.).

Split-half (Spearman Brown formula) internal consistency reliabilities for the items range from .84 to .95 for a class size of 35-49, with lower reliabilities possible for smaller class sizes, and relatively high reliabilities for larger classes (Hoyt & Lee, 2002). Validity evidence for the IDEA consists entirely of internal comparisons within the system. For example, student ratings of progress on objectives tend to be higher for objectives selected by teachers than for those the teachers did not select. In other words, even though the teachers and students make independent ratings, student ratings follow the pattern that would be predicted from teacher's ratings. Several other pieces of information are similarly internally consistent in a way that belies the validity of the ratings, but formal tests of validity have not been conducted with outside measurements (Hoyt & Lee, 2002).

The factor structure of the IDEA is partially ambiguous, as items did not load cleanly onto separate factors (some loaded onto multiple factors). Nonetheless, there appear to be three factors within faculty chosen objectives: intellectual development, professional preparation, and basic cognitive development. However, student ratings of the methods used to reach those objectives evidenced only two factors: the first included the teacher's role in transmitting knowledge, the second the student's role in acquiring knowledge (Hoyt & Lee, 2002). Nonetheless, one major advantage of the IDEA system is the ability of a teacher or administrator to compare ratings to others who taught similar courses at similarly-sized institutions. That normative information helps combat the ubiquitous halo effect seen in many other ratings by providing a frame of reference. For example, ratings might appear high when considering the range of the rating scale but average when compared to other professors of similar classes.

Considerations for Teaching Assistants

Often, teaching assistants (TAs) who are substantially involved with the running of a course want feedback on their performance. This feedback is especially important if the TAs have substantial face-to-face interaction with students and if they plan to develop their skills in a formative fashion—for example, if they want to teach a course independently. However, it is rare that the developers of SEIs consider such needs when constructing an instrument. Of the systems reviewed in this chapter, only the University of Washington program has forms specifically developed for TAs. Nonetheless, it can be perfectly appropriate for a TA to use the same form as was used for the course overall, with some caveats. It should be noted that certain items may not apply (such as test construction or planning the structure of a course). Similarly, there may be some "perceptual overlap" with how the students view the course and the TA. In other words, the students' ratings of the course overall might color those of the TA's performance as well. Statistically, one could attenuate the ratings of a TA by controlling for overlap with ratings for the instructor of the overall course. One need not be that sophisticated, however, and may simply view the TA's ratings through the lens of how students felt about the course overall.

Conclusion

There are a number of concerns to consider when selecting an SEI and even more options to weigh. Nonetheless, the most important consideration continues to be how one intends to use the SEI, as this will drive all of the other considerations. Hopefully the guidelines offered in this chapter will help in the selection of an SEI, be it one of the instruments reviewed in the chapter or one of the many others in existence. Regardless of the instrument you choose, responsible use of the instrument requires an understanding of how the instrument was developed so that you can judge how its scores might be interpreted for your particular purpose.

References

Addison, W. E. (2005, August). *The multidimensionality of effective teaching: Evidence from student evaluations.* Paper delivered at the American Psychological Association Convention, Washington, DC.

Barnes, D., Engelland, B., Matherine, C., Martin, W., Orgeron, C., Ring, J, Smith, G., & Williams, Z. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal*, *42*, 199-213.

Buskist, W., Sikorski, J., Buckley, T., & Saville, B. K. (2002). Elements of master teaching. In S. F. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 27-39). Mahwah, NJ: Erlbaum.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*, 495-518. doi: 10.1023/A:1025492407752

Chism, N. V. N. (2004). Characteristics of effective teachers in higher education: Between definitional despair and uncertainty. *Journal on Excellence in College Teaching, 15,* 5-35.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement*, *5*, 99-108.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Elton, L. (1998). Dimensions of excellence in university teaching. *International Journal for Academic Development, 3,* 3-11. doi: 10.1080/1360144980030102

Epting, L. K., Zinn, T. E., Buskist, C., & Buskist, W. (2004). Student perspectives on the distinction between ideal and typical teachers. *Teaching of Psychology*, *31*, 181-183. doi: 10.1207/s15328023top3103_5

Feeley, T. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education*, *51*, 225-236.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

Gillmore, G. (2000). *Drawing inferences about instructors: The inter-class reliability of student ratings of instruction* (Office of Educational Assessment Report 00-02). Retrieved from http://www.washington.edu/oea/pdfs/reports/OEAReport0002.pdf

Gillmore, G., & Greenwald, A. (1994). *The effects of course demands and grading leniency on student ratings of instruction* (Office of Educational Assessment Report 94-4). Retrieved from http://www.washington.edu/oea/pdfs/reports/OEAReport9404.pdf

Gump, S. E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly*, *30*, 55-68.

Hoyt, D. P. (1973). Measurement of instructional effectiveness. *Research in Higher Education*, *1*, 367-378. doi: 10.1007/BF00991670

Hoyt, D. P., Chen, Y., Pallett, W. H., & Gross, A. B. (n.d.). *Revising the IDEA system for obtaining student ratings of instructors and courses* (Technical Report 11). The IDEA Center: Manhattan, Kansas. Retrieved from http://www.theideacenter.org/sites/default/files/techreport-11.pdf

Hoyt, D. P., & Lee, E. (2002). *Basic data for the revised IDEA system* (Technical Report 12). The IDEA Center: Manhattan, Kansas. Retrieved from http://www.theideacenter.org/sites/default/files/techreport-12.pdf

Keeley, J. W., Christopher, A., & Buskist, W. (2012). Emerging evidence for excellent teaching across borders. In J. E. Groccia, M. Alsudairi, & W. Buskist (Eds.). *The handbook of college and university teaching: Global perspectives* (pp. 374-390). Thousand Oaks, CA: Sage Publications, Inc.

Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the Teacher Behavior Checklist. *Teaching of Psychology, 37,* 16-20. doi: 10.1080/00986280903426282

Keeley, J. W., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology*, *33*, 84-90. doi: 10.1207/s15328023top3302_1

Kreber, C. (2002). Teaching excellence, teaching expertise, and the scholarship of teaching. *Innovative Higher Education*, *27*, 5-23. doi: 10.1023/A:1020464222360

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mismeasurement of performance. *Economics of Education Review*, *27*, 417-428. doi: 10.1016/j.econedurev.2006.12.003

Lowman, J. (1995). *Mastering the techniques of teaching* (2nd ed.). San Francisco: Jossey-Bass.

Marsh, H. W. (1981). Students' evaluations of tertiary instruction: Testing the applicability of American surveys in an Australian setting. *Australian Journal of Education*, *25*, 177-192.

Marsh, H. W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77-95. doi: 10.1111/j.2044-8279.1982.tb02505.x

Marsh, H. W. (1982b). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology*, *74*, 264-279. doi: 10.1037/0022-0663.74.2.264

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707-754. doi: 10.1037/0022-0663.76.5.707

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of students' evaluations of college teaching. *American Educational Research Journal*, *21*, 341-366. doi: 10.2307/1162448

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181-210. doi: 10.1111/j.1745-6916.2006.00012.x.

Schaeffer, G., Epting, K., Zinn, T., & Buskist, W. (2003). Student and faculty perceptions of effective teaching: A successful replication. *Teaching of Psychology, 30,* 133-136.

Smith, D., Cook, P., & Buskist, W. (2011). An experimental analysis of the relation between assigned grades and instructor evaluations. *Teaching of Psychology*, *38*, 225-228. doi:10.1177/0098628311421317

University of Washington Office of Educational Assessment. (2005). *Course Evaluation*. Retrieved from http://www.washington.edu/oea/services/course_eval/index.html

Vulcano, B. A. (2007). Extending the generality of the qualities and behaviors constituting effective teaching. *Teaching of Psychology, 34,* 114-117. doi: 10.1080/00986280701293198

Wann, P. D. (2001, January). *Faculty and student perceptions of the behaviors of effective college teachers.* Poster presented at the National Institute for the Teaching of Psychology, St. Petersburg Beach, FL.

Contact Information

Jared W. Keeley, Ph.D.
Assistant Professor
Department of Psychology
Mississippi State University
PO Box 6161
Mississippi State, MS 39762
662-325-4799
jk593@msstate.edu

Formative Teaching Evaluations: Is Student Input Useful?

Janie H. Wilson and Rebecca G. Ryan
Georgia Southern University

Evaluations of teaching offer valuable feedback and take several forms, including self-, peer-, and student-generated assessment. In this chapter, we focus primarily on student evaluations of teaching (SETs) and examine content, timing, and format of evaluations. Finally, we discuss uses of teaching evaluations to effect change.

Why Use SETs?

By far, student evaluations comprise the bulk of teaching feedback (e.g., Jahangiri, Mucciolo, Choi, & Spielman, 2008); we rely on students to provide fair and accurate assessments of our teaching. Murray (1987) argued that SETs offer numerous benefits, including cost effectiveness and low sampling error. Perhaps most importantly, SETs provide reliable and valid indications of teacher performance. See Keeley, this volume, addresses how to select an SET.

In general, students attend class. As a result, teachers do not expend time and effort locating a group to offer feedback. In addition, students do not require payment to evaluate their teachers. Finally, SETs are cost effective because they produce data from a large source rather than a single peer evaluator or outside observer. In short, student evaluations offer a wealth of information at little cost.

With a large sample of evaluations, teachers should feel less concerned about sampling error. Students provide feedback based on the entirety of the instructor's performance in the classroom. They know how an instructor teaches on a consistent basis. Even if outside observers attend the class at randomly scheduled times, they can only base their evaluations on a small sample of the instructor's overall performance. If scheduled in advance, the anticipation of outside observers may increase planning time and create a positive bias. Whether scheduled or unscheduled, the presence of observers will likely increase stress and either positively or negatively influence the instructor's performance during that class session, again resulting in a nonrepresentative experience. In fact, the presence of an outside observer may alter student behaviors as well, causing a social climate that might not otherwise exist. Ismail, Buskist, and Groccia, this volume, address effective strategies for conducting peer review of teaching. The bottom line, however, is that peer review cannot replace SETs because students evaluate consistent performance in the classroom and assess a representative sample of teaching performance.

Reliability and Validity of SETs

A wealth of evaluation data based on representative slices of teaching holds little value if the SETs themselves fail to offer reliable and valid measures. Researchers have investigated both inter-rater and test-retest reliability of student ratings and found support for both (e.g., Marsh & Roche, 1997; Murray, 1983). However, correlations between any two students in a class are about .20; class-level reliability increases dramatically for larger groups, to approximately .90 for 25 students (Marsh & Roche, 1997). SETs are also valid based on positive correlations between SETs and teacher self-assessments (Keeley, this volume, Marsh & Roche, 1997), although correlations are generally moderate. Significant agreement was found between student ratings and teacher ratings on each dimension of the Student's Evaluation of Educational Quality (SEEQ) assessment. These dimensions include; "Learning/Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Examinations/Grading, Assignment/ Readings, and Workload/Difficulty" (Marsh & Roche, 1997).

Further, student ratings of overall teaching effectiveness yielded a correlation of .32 with teacher self-evaluations. Correlations between SETs and teacher self-evaluations for specific teacher qualities (aforementioned dimensions) were higher, ranging from .45-.49. (For a thorough review of SET reliability and validity, see Marsh & Roche, 1997.) Significant relationships between related constructs indicate validity even when overlap also illustrates that different constructs exist. Thus, SETs converge on quality of teaching based on measures of reliability and validity, avoid sampling error based on small slices of teaching performance, and are cost effective. In general, SET values fluctuate slightly or none based on class size, grades, and workload, to name a few (e.g., Remedios & Lieberman, 2008; Marsh & Roche, 1997).

If student evaluations provide useful information, how soon should teachers gather that information? Evaluations take the form of either summative or formative. The former offers end-of-term assessments of teaching and may or may not cause a teacher to adjust the course in a subsequent term. Problems reported by students merely reflect a history that cannot be changed. In effect, feedback comes too late to be useful to the class. As an alternative, formative assessments provide feedback prior to the end of the term, allowing teachers to alter courses as needed and improve the teaching-learning experience. Berk (2005) proposed that formative evaluations reflect a desire to improve teaching, whereas summative evaluations are used by administrators to decide promotion, tenure, and paychecks. For the purpose of this chapter, we focus on formative student evaluations of teaching based on the idea that teachers use evaluations to improve their teaching. We discuss the content of SETs, evaluation procedures, and teacher uses of formative assessments.

What to Evaluate

Teaching evaluations measure quality of teaching (Remedios & Lieberman, 2008). Wilson (1986) reported five factors that represent good teaching: (1) organization/clarity, (2) dynamism/enthusiasm (3) analysis/synthesis (e.g., contrasts theories and discusses viewpoints other than his/her own), (4) teacher-group interaction (e.g., encourages class discussion), and (5) teacher-student interaction (e.g., genuinely interested in students). Marsh and Hocevar (1984) proposed nine evaluation dimensions. We can see a clear overlap with Wilson on several: (1) organization, (2) enthusiasm, (3) group interaction, and (4) individual rapport. Wilson's "analysis/synthesis" category perhaps can expand into (5) learning/value, (6) breadth of coverage, (7) examinations, (8) assignments, and (9) workload/course difficulty. Organization, clarity, and the nuts and bolts of teaching a class (e.g., assignments and workload) seem to get at the heart of competent teaching, but teachers should not ignore the importance of enthusiasm and rapport. In fact, research on student-teacher rapport represents a relatively recent addition to SETs, and research on classroom rapport flourishes. Marsh (1984) makes the case that a good SET instrument must reflect the multidimensional nature of teaching and offers a review of several evaluations with support through factor analysis and a great deal of overlap among key factors (see p. 711 of Marsh, 1984, for several scales).

Additional SET measures not likely to catch on include instructor easiness (low requirements of student performance), clothing, attractiveness, and even hotness. Teachers often report random comments from students concerning such inappropriate topics, and ignoring these comments remains the only realistic option. In fact, students may benefit from knowing what type of formative comments might improve the class versus those that make the teacher uncomfortable. As a caution, however, some personal remarks provide valuable information. During college, the first author attempted to learn from a professor who self-mutilated during class; he repeatedly scratched the skin from his finger until it bled. He probably assumed no one noticed. In fact, perhaps he failed to notice his own behavior. But

formative feedback from students should allow him to stop scratching and refocus student attention on the course material.

When to Evaluate

Administering formative student evaluations during early-to-mid semester provides input on teaching quality and allows enough time for instructors to modify their teaching in areas where they are rated poorly. Basic procedures enhance the data. For example, the majority of the class should provide assessment. Although teachers may feel tempted to collect SETs when only the most dedicated students come to class, the goal includes learning perceptions of both strong and weak students. Acquiring a full complement of SETs may translate into not administering evaluations on the last day of classes before a holiday break, over homecoming weekend, the class period after a major test, or even on a Friday.

Instructors should also schedule their evaluations at times when students are least likely to provide biased feedback. For example, instructors should not collect evaluations after awarding extra credit, postponing a deadline, offering an unusually engaging activity, or handing out candy. Dickey and Pearson (2005) discussed the power of the recency effect when asking students to provide evaluations immediately following a positive class experience. Essentially, instructors should not ask for SETs following an unusual event; after all, the goal is to obtain a truly representative sample of consistent performance. Simpson and Siguaw (2000) offered a fascinating list of faculty practices meant to enhance SETs, including grading leniency offered by 23.6% of their sample! Certainly, instructors should not intentionally attempt to bias SETs in the positive direction. By the same token, SETs should not be collected during particularly negative times, such as the anniversary of 9/11 or following a stressful discussion of a controversial topic (unless that is the class norm). Such details of administering evaluations may not be intuitive. While an undergraduate, the second author's professor said he was going to add a substantial number of points to a previous exam and urged students to remember that favor when completing evaluations, which he distributed the following day.

Where to Evaluate

Traditionally, teachers collect SETs during a class period from physically-present students, and the majority of research focuses on the classroom context. However, due to enhanced technology and readily available internet access, students may evaluate their teacher online. Several variations occur, including

- Optional course evaluations online
- Required course evaluations (e.g., grade withheld until student completes evaluation)
- Institution-sanctioned online evaluations for an entire campus
- Student-motivated online evaluations such as RateMyProfessors.com
- Teacher requests for students to evaluate on RateMyProfessors.com

Most online evaluations offer students a chance to provide feedback about professors at a location and time of their choosing, reduce time constraints, and reduce potential influence from the professor (Anderson, Cain, & Bird, 2005). Online evaluations also decrease the problem of students who do not get the chance to complete in-class evaluations if they miss class during evaluation day. (See Adams, this volume, for issues related to online SETs.)

Online SETs are gaining popularity, but if they remain optional, response rates suffer. For example, Nowell, Gale, and Handley (2010) reported an online response rate of 28.1% compared to an in-class

rate of 72.2%. In addition, Nowell and colleagues estimated class ratings .69 points lower online; individual instructors received approximately .81 points lower online. A nonrepresentative (disgruntled) online sample may explain lower ratings. Conversely, Donavan, Mader, and Shinsky (2006) specifically requested all students to provide online evaluations and found similar responses across the two methods. Interestingly, students evaluating online offered more open-ended responses than students evaluating in the classroom. Urging from the professor may have created a more representative sample of evaluations, and more free-response feedback increases the potential value of online SETs.

Of course, requiring online SETs of students allows for a healthy sample of evaluations. Davidson and Price (2009) suggested that universities create their own rating websites for students. Such websites could include items to assess teaching effectiveness, course design, workload, grading, preparation for class, and other items that would evaluate teaching. When institutions require students to complete online evaluations prior to obtaining course grades, response rates soar.

Websites such as RateMyProfessors.com cater to students and provide a forum for college students to evaluate their professors. Although teachers may not like some of the items (e.g., hotness), Davidson and Price (2009) reported that 80% of a college-student sample reported using the site to learn more about a professor. A separate study indicated 83% of students used the site (Brown, Baille, & Fraser, 2009). Legg and Wilson (2010) created more representative samples (similar to in-class evaluations) on RateMyProfessors.com by encouraging all students to use the site.

A more recent option for quick evaluations is Twitter. Twitter is an electronic method of brief communication; text-based posts can be up to 140 characters and are considered personal "updates" send out to a group of people who have signed up to "follow" a specific person. Steiger and Burger (2010) suggested that Twitter allows students to provide weekly formative assessments at little cost and without taking up valuable course time. Comparisons with formative in-class evaluations showed no differences. The authors caution that such frequent assessment (and the nature of Twitter) requires only a few brief items. They used, "How did you like the course today?" and "How interesting was the course today?" with a rating scale from 1-9. In addition, students could choose to report "What was good today?" and "What was bad today?" as well as offer general comments.

Uses of Formative Teaching Evaluations

In prior sections, we presented information about the content of SETs, suggestions for when to administer evaluations, and options for the context of SETs. Now we turn to the uses of evaluations. Do they improve teaching? If so, how?

The majority of student evaluations are collected at the end of the term as summative assessment. Overall, do summative SETs improve teaching? Kember, Leung, and Kwan (2002) suggested that they do not. Several years of SET data collapsed across departments failed to reveal any teaching improvements as measured by subsequent student evaluations. In fact, three of the four changes that did emerge revealed a decrease in teaching quality. Similarly, Marsh (2007) studied data from individual faculty members across 13 years of teaching and found no improvements, even though SETs provided feedback year after year. We conclude that simply collecting SETs as a matter of rote falls short of enhancing teaching.

Cohen (1980) included 22 research studies in a meta-analysis of research investigating the usefulness of feedback from mid-semester student evaluations. Of the studies reviewed, Cohen reported that 10 found statistically significant differences in end-of-semester student evaluations. Instructors who were

given feedback at mid-semester had higher end-of-semester ratings of skill, rapport, structure, difficulty, interaction, feedback, and overall teaching effectiveness compared to those who did not receive mid-semester feedback. Although only 10 significant differences emerged, Cohen found means in the predicted direction for 20 of the 22 studies.

Similarly, Menges and Brinko (1986) conducted a meta-analysis of research investigating the usefulness of mid-semester student feedback. Of a total of 30 studies, they found that 10 revealed statistically significant differences between those who received mid-semester feedback compared to those who did not. In fact, ratings for those who received feedback were higher than 67 percent of the ratings for those who did not receive feedback. We should also note that Menges and Brinko further examined studies by type of mid-term feedback, including student ratings, consultation, and feedback from additional sources like self-evaluation, peer-evaluation, or peer-group feedback. When examined based on type of feedback, they indeed found a beneficial effect of SETs alone; however, the effect was small. Overall, those who received feedback enjoyed average end-of-semester ratings that were higher than 59 percent of the control-group's ratings. When SETs and consultation comprised mid-term evaluations, those who received student and consultant feedback averaged end-of-semester ratings higher than 86 percent of the control-group's ratings. These data suggest that SETs remain useful, but input from consultants clearly enhances assessment.

Wilson (1986) offered a step-by-step process for effective use of consultants. First, summative evaluations were collected at the end of the term; in effect, they represented formative assessment to improve subsequent classes. Second, consultants briefly discussed SETs with the faculty member and asked for areas in which the instructor would like to improve. Third, the instructor scheduled an indepth consultation to be held between two and four weeks prior to the beginning of the subsequent class. For that meeting, the consultant brought a few low-scoring items on the SETs as well as reviewed open-ended student comments for useful information. Consultants discussed weak areas and offered ideas for improvement. A few days following the meeting, the consultant contacted the teacher by letter, providing a summary of the meeting and reiterating ideas for improvement. A phone call during the term reinforced the consultant's willingness to help. SETs were collected again at the end of the term, and some instructors chose to repeat the entire process and collect SETs at the end of a third term. Wilson reported that 52% of faculty members involved in the consultation process enjoyed significant improvement in teaching effectiveness. We should note that Wilson utilized consultants with no administrative power over the instructors involved. Further, Brinko (1993) argued that an effective consultant maintains confidentiality and is authentic, respectful, supportive, empathic, and non-judgmental. Though consultations do require additional time, effort, and perhaps cost, research does support an improvement beyond using student feedback alone. (See Ismail et al., this volume for a discussion of how to conduct effective peer review of teaching.)

We might feel compelled to explain more positive evaluations by improvements to teaching, and indeed that may be the case. Alternative explanations paint a less-rosy picture: For example, students might purge their negative perceptions at mid-term and focus on positive perceptions at the end of the term. As a third potential explanation, students may simply believe that their instructor will change based on feedback and seek evidence of improvement that may or may not exist. Rather than see these alternative explanations as problematic, we prefer to focus on the positive social aspect at work (see below).

Students appreciate being asked to provide feedback about the quality of their instructors' teaching. This practice communicates to students that their opinions are valued, and they provide a useful source

of assessment. Students, particularly the current generation of students, appreciate being given a voice and the opportunity to express themselves (Twenge, 2006). This practice may be especially helpful in large classes to reduce perceived anonymity and disconnect from the instructor. But even in small classes, students want to be heard.

Thus, SETs build a sense of ownership and community that is beneficial to teaching and learning. Asking students for input accentuates the social nature of teaching. As an added benefit, collecting SETs builds student-teacher rapport, which positively correlates with student attitudes toward the instructor and course, student motivation, perceived learning, and even course grades (Wilson, Ryan, & Pugh, 2010). In fact, given the numerous links between rapport and positive student outcomes, teachers may want to include an assessment of rapport in their SET forms (Ryan, Wilson, & Pugh 2011; Wilson et al., 2010). If rapport is lacking, teachers can address that specific problem by focusing on building more positive relationships with students.

Conclusions

Formative evaluations provide an opportunity for students to express their opinions. Students generously offer positive evaluations, and often they share valuable feedback that allows a teacher to adjust the course in some way. In the end, students want to be heard; they want to know the teacher is listening.

Formative evaluations from students require faculty reflection in order to foster change. When teachers utilize only surface-level processing of student feedback, responses illustrate a reactive style (Winchester & Winchester, 2011). That is, merely reading evaluations, and perhaps summarizing them for an annual review, fails to elicit useful change. On the other hand, thoughtful, deliberate processing of student feedback leads to a proactive style; teachers consider avenues for change based on careful consideration of student perceptions. Certainly, teachers may decide not to alter some aspect of a course, but negative student reactions may lead teachers to clarify course practices to students. At the very least, deep reflection of a teaching practice solidifies precise reasons for the practice.

As teachers, we learn from student evaluations, and formative assessment paints a vivid picture of student perceptions. Although SETs remain the most widely accepted and utilized form of teacher and course assessment, we can benefit from peer evaluation of teaching as well as self-reflection (see Schafer, Yost Hammer, & Bernsten, this volume). (For a comprehensive literature review of the potential variables influencing SETs, peer evaluations, and self-assessment, see Tables 1-3 in Jahangiri et al., 2008). “Triangulation” refers to SETs, peer evaluations, and self-reflection converging on the teaching process (Appling, Naumann, & Berk (2001). Berk (2005) also called this approach a “unified conceptualization” of teacher effectiveness and proposed 12 sources for evaluation of teaching. In addition to SETs, peer input, and self-reflection, Berk added videos, student interviews, teaching scholarship and awards, a teaching portfolio, learning-outcome measures such as student grades, and alumni, employer, and administrator ratings. Triangulation pinpoints areas ripe for improvement. As an added benefit, several types of assessment reveal our strengths. We gain the opportunity to weed out bad habits and fine-tune useful approaches to teaching.

Given the potential advantages of SETs, we recommend that all instructors collect formative teaching evaluations. However, teachers must be careful to collect SETs in an ethical manner rather than attempt to maximize positive ratings. Peer evaluations and self-reflection promise to paint a clearer picture of teaching effectiveness than SETs alone. Finally, consulting with a supportive colleague allows student evaluations to foster long-term change in teaching.

References

Anderson, H. M., Cain, J., & Bird, E. (2005). Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education, 69,* 34-43. doi: 10.5688/aj690105

Appling, S. E., Naumann, P. L., & Berk, R. A. (2001). Using a faculty evaluation triad to achieve evidence-based teaching. *Nurse Health Care Perspective, 22*, 247-251. doi: 10.1043/1094-2831(2001)022<0247:UAFETT>2.0.CO;2

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. International *Journal of Teaching and Learning in Higher Education, 17*, 48-62.

Brinko, K. T. (1993). The practice of giving feedback to improve teaching: What is effective? *Journal of Higher Education, 64*, 574-593. doi: 10.2307/2959994

Brown, M. J., Baillie, M., & Fraser, S. (2009). Rating RateMyProfessors.com: A comparison of online and official student evaluations of teaching. *College Teaching, 57*, 89-92. doi: 10.3200/CTCH.57.2.89-92

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education*, *13*, 321-341. doi: 10.1007/BF00976252

Davidson, E., & Price, J. (2009). How do we rate? An evaluation of online student evaluations. *Assessment & Evaluation in Higher Education, 34*, 51-65. doi: 10.1080/02602930801895695

Dickey, D., & Pearson, C. (2005). Recency effect in college student course evaluations. *Practical Assessment, Research & Evaluation, 10*, 1-10.

Donovan, J., Mader, C. E., & Shinsky, J. (2006). Constructive student feedback: Online vs. traditional course evaluations. *Journal of Interactive Online Learning, 5,* 283-296.

Jahangiri, L., Mucciolo, T. W., Choi, M., & Spielman, A. I. (2008). Assessment of teaching effectiveness is U.S. dental schools and the value of triangulation. *Journal of Dental Education, 72*, 707-718.

Kember, D., Leung, D. Y. P., & Kwan, K. P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education, 27*, 411-425. doi: 10.1080/0260293022000009294

Legg, A. M., & Wilson, J. H. (2010). RateMyProfessors.com Offers Biased Evaluations. *Assessment & Evaluation in Higher Education*. doi:10.1080/02602938.2010.507299

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity potential biases, and utility. *Journal of Educational Psychology*, *76*, 707-754. doi: 10.1037/0022-0663.76.5.707

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*, 253-388. doi: 10.1016/0883-0355(87)90001-2

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*, 775-790. doi: 10.1037/0022-0663.99.4.775

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal, 21*, 341-366. doi: 10.3102/00028312021002341

Marsh, H. W., & Roche, L. A. (1997). Making student evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187-1197. doi: 10.1037/0003-066X.52.11.1187

Menges, R. J., & Brinko, K. T. (1986). *Effects of student evaluation feedback: A meta-analysis of higher education research.* Paper presented at the 70th meeting of the American Educational Research Association, San Francisco, CA. doi: 10.3102/00346543074002215

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, *75*, 138-149. doi: 10.1037/0022-0663.75.1.138

Murray, H. G. (1987). Acquiring student feedback that improves instruction. *New Directions for Teaching and Learning*, *32*, 85-96. doi:10.1002/tl.37219873210

Nowell, J. B., Gale, L. R., & Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment and Evaluation in Higher Education, 35*, 463-475. doi: 10.1080/02602930902862875

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal, 34*, 91-115. doi: 10.1080/01411920701492043

Ryan, R. G., Wilson, J. H. & Pugh, J. L. (2011). Psychometric characteristics of the professor-student rapport scale. *Teaching of Psychology*, *38*, 135-141. doi: 10.1177/0098628311411894

Simpson, P. M., & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education, 22*, 199-213. doi: 10.1177/0273475300223004

Steiger, S., & Burger, C. (2010). Let's go formative: Continuous student ratings on Web 2.0 application Twitter. *Cyberpsychology, Behavior, and Social Networking, 13*, 163-167.

Twenge, J. M. (2006). *Generation Me*. Simon & Schuster, Inc.

Wilson, J. H., Ryan, R. G., & Pugh, J. L. (2010). Professor-Student Rapport Scale Predicts Student Outcomes. Teaching of Psychology, 37, 246-251. doi:10.1080/00986283.2010.510976

Wilson, R. C. (1986). Improving faculty teaching: Effective Use of Student Evaluations and Consultants. Journal of Higher Education, 57, 196-211. doi: 10.2307/1981481

Winchester, T. M. & Winchester, M. (2011). Exploring the impact of faculty reflection on weekly student evaluations of teaching. International Journal for Academic Development, 16, 119-131. doi: 10.1080/1360144X.2011.568679

Contact Information

Janie H. Wilson, jhwilson@georgiasouthern.edu, 912-478-5580
Rebecca G. Ryan, rgryan@georgiasouthern.edu, 912-478-5447

Using Student Feedback as *One* Measure of Faculty Teaching Effectiveness

Maureen A. McCarthy
Kennesaw State University

Over the course of the last decade, the public, along with many disciplinary organizations have increased the pressure on institutions of higher learning to evaluate student learning and instructional effectiveness (Diamond, 2008; Dunn, McCarthy, Baker, & Halonen, 2011; Dunn, McCarthy, Baker, Halonen, & Hill, 2007). College and university administrators, legislators, families, and students themselves, are among the stakeholders who want assurances that educators are delivering on their promise to educate the next generation. However, documenting student learning, together with evaluating faculty teaching effectiveness, continues to be a challenge for administrators and faculty alike. Not only are institutions required to provide assurances of learning for the stakeholders, but there is an expectation that data will be used to improve instruction (Pusateri, in press).

Both learning and teaching are multidimensional constructs; therefore, an evaluation of teaching should involve a multimethod approach (Marsh & Roche, 1997), with Student Evaluations of Teachers (SETs) comprising just one of the measures. Although student evaluations (i.e., student opinions) comprise one measure of faculty effectiveness, they have been challenged as a useful method for evaluating teaching excellence (Ballard, Rearden, & Nelson, 1976; Beyers, 2008; Buskist, 2006; Eckert & Dabrowski, 2010; Germaine & Scandura, 2005; McKeachie, 1997). Yet, student ratings do provide an important perspective that can be used to help faculty reflect on their pedagogy (Greenwald, 2006); see also Keeley, this volume and Wilson and Ryan, this volume) and student feedback can be used, if obtained and interpreted carefully, by administrators as one measure to evaluate teaching. In fact, SETs are widely used when conducting annual reviews, tenure and promotion decisions, and merit-based decisions. In this chapter, I will briefly review the research on using SETs (for a comprehensive review of instruments see Keeley this volume) to evaluate faculty and provide recommendations for how SETs can be used to both evaluate and improve instruction. I will also consider how just one measure of teaching effectiveness, SETs, can be used collaboratively by administrators and faculty to evaluate progress toward good teaching.

Review of Student Evaluations of Teaching

Although the pressure to evaluate student learning and student satisfaction has recently increased, researchers have long addressed ways to do so. Indeed, the process of developing reliable and valid instruments to measure student satisfaction began over four decades ago. (Keeley, this volume; Marsh, 1982; Keeley, Smith, & Buskist, 2006; Keeley, Furr, & Buskist, 2010). SETs were initially developed for two purposes – evaluating teaching and providing faculty with feedback. In perhaps the most comprehensive attempt to create a reliable and valid SET, Marsh (1982) examined the Student Evaluation of Educational Quality (SEEQ) instrument and found support for nine factors (Marsh & Hocevar, 1984; Marsh, 1987). Five of the factors (organization, breadth of coverage, grading, assignments, and workload) were characterized as professional competency and communication skills, and the remaining four factors (learning value, enthusiasm, rapport, group interaction) reflect affective attributes that are thought to be influential in learning. More recently, Keeley, Smith, and Buskist (2010) found that the Teacher Behavior Checklist provides a similar two factor structure (i.e., professional competency and affective attributes). In addition, regardless of the measure used, two important elements continue to comprise the basic elements of SETs – quantitative and qualitative measures.

Both the SEEQ and the Teacher Behavior Checklist (TBC; Keeley, Smith, & Buskist, 2006) provide quantitative multidimensional data reflecting students perceptions of teacher performance. Quantitative data provide important comparative information that comprises one dimension of a multimethod approach to evaluating instructor effectiveness and these measures (i.e., SEEQ, TBC) typically provide a quantitative measure of professional competency and communication skills that are essential components of good instruction. However, appropriate use of the data has often been ineffective and circumspect. For example, individual item means are frequently used to evaluate faculty performance relative to department, college, or university averages (McKeachie, 1997; Smith, 2008). Using an average score, without considering the standard deviation, standard error, outliers, frequencies, class size, or situational context may yield inaccurate interpretations. Not only are interpretations potentially unreliable (Marsh, 1982; McKeachie, 1997), but using means or medians in the absence of accompanying descriptive data can be potentially misleading. Moreover, there is a potential for bias, especially when factors such as instructor gender and discipline interact(see Basow & Martin, this volume). In the section that follows, I will provide recommendations for using quantitative feedback effectively.

Quantitative Data

How should we use the quantitative information from SETs to evaluate faculty effectiveness? The first step in using quantitative data from SETs is to examine the instrument itself. In all likelihood a locally-developed instrument includes some measure of professional competency and communication skills comparable to those instruments developed by Marsh (1987) and Buskist et al?. Again, these criteria are well suited as objective measures that offer students the opportunity to provide feedback and insightinto their classroom experiences. For example, students can offer an important perspective on how the exams related to course content and whether lectures were well organized. If students reported that the course content and exam were not aligned, then it is possible that the faculty member is not performing well in this area of professional competency – linking instruction to assessment. However, as with all feedback from students, responses to individual items should be interpreted in context and with caution. Similarly, if some students reported that the exam and course content were not aligned, it is important to more carefully examine responses to this item. For example, did only a small percentage of students report a lack of alignment? Did qualitative responses also reflect a similar trend? Insight into this feedback can also be derived by asking the faculty member to examine the course for potential inconsistencies between instruction and assessment.

Evaluation of teaching techniques using quantitative measures provides both the faculty member and evaluator with useful information about specific pedagogical competencies. These criteria (e.g., organization, breadth of coverage, workload, and assignments) can be interpreted using quantitative data as a primacy measure and, as discussed later, qualitative measures as a secondary measure. Quantitative measures of affective dimensions (e.g., rapport, approachable, respectful) of teaching also provide an evaluator with an important indicator of teaching effectiveness. The SEEQ (Marsh, 1984) and the Teacher Behavior Checklist (Buskist et al., 2002; Keeley, et al., 2006) contain both quantitative measures of professional competency and affective qualities.

In general, each SET item should be evaluated for potential merit with regard to expectations for faculty within a department or college. Evaluation of each item on the SET is particularly important with the increased number of courses offered in an online format, as some items may no longer be directly applicable (see Drouin, this volume). For example, if the instrument asks students to indicate whether the instructor met class regularly and on time, this item may no longer be appropriate for an online class format. Although it may not be possible to alter a university-wide instrument to meet department

needs, it is possible to carefully interpret each item relative to department expectations. Similarly, it is possible to emphasize specific items that may be of more importance to the department or individual faculty member.

Evaluating professional competencies

The literature examining the use of SETs spans approximately forty years and almost every institution of higher learning uses quantitative SETs as a method of obtaining feedback about instruction (Greenwald, 1997). Indeed, as noted, students can provide useful information about their experiences from a student perspective, particularly on those instructional dimensions that are considered professional competencies (e.g., organization, grading). For example, let's consider faculty responsiveness as one dimension of professional competency that can be reflected in student feedback. Quantitative student evaluations often include an item asking students to rate faculty responsiveness (e.g., responses to email, grading). But how is faculty responsiveness defined? Zinn (2008) and Price (2009) recently documented the trend for the millennial student to expect responses in a matter of minutes, rather than hours (or days). Are faculty responding to questions in a in a timely fashion? Are they grading papers in a reasonable amount of time? Providing timely and thoughtful feedback is an important obligation of faculty members; however expectations for the speed of responses have continued to shrink, so in this example, responsiveness should be defined before interpretations are made.

Another interpretation of timely feedback is reflected in the Just-In-Time-Teaching (JITT) technique. Saville (2008) found that providing students with feedback at a critical juncture (i.e., Just-In-Time-Teaching) is an important and effective pedagogical technique that enhances student learning. Thus, timely feedback is one measure of professional competency that reflects sound pedagogical practice, it can be measured using SETs, and it is a useful quantitative measure for evaluating faculty performance. Individual SET items can provide insight into performance on specific competencies, but it is also important to consider these items within the larger context.

Upon determining that the SET contains items that reflect the competencies valued by the department, and that the items are relevant to the course, quantitative criteria can be used to establish thresholds for adequate performance. How should such criteria be established? Departments often use a normative approach, such that means are calculated for individual items, along with comparative values across courses, or types of courses (Smith, 2008). However, although extreme mean values provide a gross measure of effectiveness for each of the items, means are frequently misinterpreted. In many instances at least one student in a course is disgruntled and the single outlier will obscure an otherwise high mean value, particularly in small classes. Similarly, a single positive rating may obscure otherwise problematic ratings. To avoid the mistake of inappropriately using a mean value, it is tempting to use the median value as a more accurate measure of relative efficacy. Although the median may reflect a more accurate measure of central tendency, it is important to consider also measures of variability, along with the response rate. All too often, low response rates or high drop rates are not factored into the normative evaluation, so response rates should be considered as key criterion that may not otherwise be reflected in an oversimplified interpretation of quantitative data. It is important to note that response rates typically are lower when SETs are obtained online, rather than through traditional administration (see Adams, this volume).

In addition to specific criteria reflected in individual items, evaluators must consider the distribution of responses to each item. So if, as with our previous example, a high proportion of the students report that the exams are aligned with instruction, then one can reasonably conclude that the faculty member is doing a good job of meeting the criterion. If, however, a large proportion of students report a

disconnect between teaching and assessments, then the evaluator should examine this criterion more closely. The evaluator (e.g., department chair) may want to refer to the faculty members' self-evaluation and analysis of item for clarification about why students report inconsistency between instruction and assessment. Such evaluations can take the form of peer assessment (see Ismail, Buskist, & Groccia, this volume) or a portfolio (see Schafer, Yost Hammer, & Berntsten, this volume).

A second, and perhaps more equitable method of using quantitative data for evaluating the professional competency dimension, is to consider a criterion-based approach. For example, on a scale of 1 to 5, is there a specific threshold that is considered acceptable and meets the criterion for meeting student needs? What role does variability play in the criterion-based analysis of quantitative data? Individual SET items can be used to ensure that faculty have met the minimum absolute criterion. A caution - ranking faculty within a department is not psychometrically sound, as the data are reduced to an ordinal scale, nor does a ranking provide useful insight into trends that may be present across items on the SETs. Establishing an absolute criterion should reflect a value that is independent of the average score. Moreover, as I indicted earlier, a mean ranking for an item may be influenced by outliers. So, all faculty members in a department may be doing very well, exceeding the specified criterion; if so, minimal differences on individual items won't reflect important meaningful distinctions in teaching effectiveness. Similarly, small differences in quantitative values may obscure serious problems that are reflected by qualitative measures. Ultimately, as I will discuss later, it is important to consider these quantitative items within the context of qualitative measures of faculty effectiveness.

Evaluating affective dimensions

SETs usually contain items reflecting affective dimensions (e.g., enthusiasm, rapport) in addition to pedagogical competencies. Marsh (1987) identified both pedagogical competency and affective dimensions of the SEEQ, and the Teacher Behavior Checklist (Buskist, Sikorski, Buckley, & Saville, 2002) is comprised almost exclusively of items reflecting affective constructs. These affective behaviors are arguably important, but interpretations of these dimensions are understandably more complicated. For example, demonstrating respect for students is absolutely critical for learning, and it is easy to defend the importance of a high score this affective dimension. However, evaluating a dimension such as the approachability of a faculty member using a rating scale is a bit more complicated. What does a high score on this item mean? It is entirely possible that a highly professional instructor (e.g., one who establishes clear boundaries) may be perceived as less approachable than an instructor how is perhaps too friendly with students. Ratings of approachability are also confounded with gender (Bachen, Mcloughlin, & Garcia, 1999; Basow & Martin, this volume) and other factors (see Keeley, this volume and Wilson & Ryan, this volume). So, although affective dimensions of SETs are important, interpreting the data based on these items can be complicated.

Indeed, some believe (Buskist, 2002; Myers, 2005) that affective behaviors can be an equally important determinant of student success. The Teacher Behavior Checklist (Keeley et al., 2006) includes a more detailed set of items that can be used to measure the affective dimension of teaching effectiveness. For example, Keeley et al. offer a description of instructor understanding that includes accepting legitimate excuses for missing class. This approach to quantitatively measuring affective behaviors provides a more accurate explanation of specific teacher behavior, and multiple questions might provide richer insight into how a teacher balances rigor with understanding. So, it may be necessary to examine two affective items simultaneously. In other words, faculty may retain high standards for student achievement, while at the same time demonstrating encouragement for student learning. Thus, the quantitative measures of affective dimensions provide an initial barometer of faculty behaviors. But again, quantitative items should include careful analysis measures of central tendency, normative versus criterion-referenced

approaches, and frequency analysis along. In addition to examining individual items, some colleges and universities produce an overall quantitative measure for the SETs that reflect the average ratings across items. This summative measure not only obscures a potential problem on a single dimension, but may also be inaccurate due to the instructor demographic variables course content. The quantitative data should also be interpreted along with qualitative measures as discussed next.

Qualitative Measures

In addition to the quantitative data that are usually in ample supply, SETs usually contain at least one measure of qualitative feedback (e.g., open-ended question) that is general in nature. This qualitative feedback is usually solicited through open-ended questions designed to elicit constructive feedback that individual faculty can use to improve their teaching. However, interpreting qualitative feedback, often reflecting the affective component of teaching, is necessarily more complex. Faculty and administrators alike often underutilize qualitative feedback and fail to detect important trends across courses or semesters. Therefore it is also important to consider how qualitative data can be used to inform faculty and provide a measure of instructional effectiveness.

How should qualitative data be used in the evaluation process? If, for example, students provide suggestions for how a course might be improved, a range of responses usually result. A parsimonious solution to this reality is not only elusive, but fails to capture the multidimensional nature of faculty behaviors and competencies. A first step may be to examine the qualitative data for trends. For example, if a large number of students report that the instructor did not use the required textbook, then it seems appropriate to reconsider requiring the text. Similarly, if only one student reports that he/she believed the workload was too high for the course, then this comment can be easily dismissed. Much like the quantitative data, it is important to look for trends that exist within and across semesters. A single comment in one semester may not be enough to raise concerns, but similar comments across semesters might be important to track, particularly in the case of tenure and promotion decisions. As with quantitative data, these qualitative data should be examined in connection with the faculty members' self-assessment.

Qualitative responses often help to clarify the quantitative responses provided on the affective dimensions. For example, the Teacher Behavior Checklist (Keeley, et al., 2006) includes an item assessing whether the instructor using a creative and interesting presentation style Although the instructor might be rated highly on this item, students might qualitatively report that the instructor tells so many interesting stories that content relevant instruction suffers. In other words, although it is important to actively engage students, and creative anecdotes might be interesting, it is possible to score highly on this item while at the same time introducing an impediment to learning that can only be identified using the qualitative responses of SETs. In sum, the multi-method analysis provides a more complete picture of affective behaviors.

The qualitative measures can also provide important insight into specific behaviors that might not otherwise be considered using a purely quantitative measure. For example, students might report that an instructor routinely engages in rigid grading practices designed to penalize, rather than encourage student learning. If we rely only on the quantitative items, then a faculty member might be rated highly on the item that asks if grading criteria are clear. However, in this example, clear criteria may be so rigid that students attend to insignificant details rather than the more important pedagogical goal of instilling critical thinking. Therefore, a single item from a typical SET may not tell the full story. Instead, a complete analysis requires qualitative measures.

As another example, if students report that they are not receiving enough guidance for a writing assignment, the responses must be considered within the context of the overall course. Providing students with adequate guidance for writing may be appropriately directed with a well-developed rubric that is provided to students. Yet extremely detailed, but marginally important criteria (e.g., losing one point for every error in APA format) may restrict students from developing the ability to write well. The rubric that a faculty member uses to evaluate an assignment could be part of a teaching portfolio; if so, evaluators could consider the clarity of the rubric in conjunction with the SET responses. Qualitative feedback derived from SETs must be carefully examined for trends that might reflect a general approach that may or may not be beneficial to student learning. For example, if a faculty member emphasizes superficial learning (e.g., format) that is easy to grade, rather than critical thinking, ratings of clarity may artificially produce lowing ratings for the teacher emphasizing a more comprehensive teaching approach. Thus, qualitative, open-ended items provide a rich source of information that can help to inform the evaluator about the overall classroom experience.

Using SETs in the Evaluation Process

Although SETs are not necessarily a measure of student learning, the combined use of quantitative and qualitative measures provides important insight into teaching effectiveness. Using a multi-method approach that includes both types of data from SETs, along with a faculty self-evaluation and pedagogical materials can provide useful information that will aid in evaluating faculty effectiveness. To obtain a complete understanding of faculty effectiveness, it is important to consider these measures within the context of the faculty members' general approach to teaching. Without an understanding of the faculty member's approach, it is possible that data may be misunderstood.

How should SETs be interpreted? First, it is important to consider situational context. Although a teaching philosophy may not always be explicitly stated, faculty members usually have a working philosophy that is transmitted to students informally. If faculty have not articulated their teaching philosophy, then it may be useful to invite faculty to be more explicit about their approach. A clear philosophy statement can then be used by the faculty member and the evaluator to place SETs in context. So, if a faculty member is using a developmental approach that involves challenging students to improve over the course of a semester, then their approach may be reflected in their assignments. For example, if students write a literature review and are provided with general criteria (i.e., a grading rubric) for how the assignment will be evaluated, then the assignment and grading practices may be pedagogically sound. Students should always receive clear guidance in advance of submitting an assignment. For example, a rubric that details how an assignment will be graded provides useful guidance for the student. Yet students may report that they were not provided with adequate guidance for the assignment. In this context, a balanced rubric, one that provides guidance, but not so highly detailed that it prevents students from developing critical thinking skills, may be appropriate. In other words, a highly detailed grading rubric that emphasizes unimportant details (e.g., rigid penalties for small APA style errors), rather than more complex concepts (e.g., general writing ability) may limit the students' developmental progression. In fact, the less specific rubric may provide students with clear guidance about assignments that is consistent with the pedagogical intentions of the instructor, but at the same time, students may be uncomfortable with the challenge to develop more sophisticated writing skills. Viewing the grading rubric aids in interpreting the SETs as responses are then being considered within the context of the faculty member's teaching philosophy. Doing so may reveal that students are provided with just the right amount of guidance to help them develop their writing skills.

What is the process for using SETs in the evaluation cycle? Who should review the SETs first? For example, during the annual review process sometimes a department chair will conduct a preliminary

review of the SETs, and in other instances the faculty member reviews the SETs first. If the SETs are to be used to improve instruction, then the faculty member should review the SETs first. A thorough self-evaluation of the feedback may help an individual faculty member to improve their instruction independently. Faculty should be encouraged to annually review the SETs in the context of their teaching approach and to develop a response to the qualitative and quantitative elements of the SETs. Providing faculty members with the opportunity to review the SETs first allows the faculty member to interpret the feedback within the context of their teaching style and to provide any additional context that may have been specific to the course. Similarly, an annual review of SETs provides an opportunity for a faculty member to address a potential weakness before a pattern emerges across multiple semesters. For example, it is possible that one section of a course inexplicably doesn't work out well and the quantitative evaluations were uncharacteristically low. It may be that one disruptive student introduced a level of discomfort that simply could not be overcome. So, in this one instance the faculty member can provide the context that helps to interpret the unusual set of SETs. When a faculty member prepares the promotion and tenure portfolio, he or she can place the SETs in context. Therefore, it only makes sense to allow faculty to engage in self-reflection and context-dependent variables annually. In other words, faculty should interpret their SETs and provide ideas for how to improve (and there is always room for improvement).

It is also important to consider how SETs should be used holistically. Although it is important to engage in thoughtful review and evaluation of SETs annually, it is also important to consider trends that may persist long-term. SETs are almost always required as one component of a tenure, promotion, and post-tenure portfolios. Multiple evaluators will have an opportunity to review SETs and to consider trends across semesters. This broader perspective will allow a faculty member and evaluators to reflect on the larger, more complete set of student feedback that reflects a sustained approach to teaching. In most instances faculty will be able consider trends across semesters and reflect on how their philosophy is reflected in the feedback they receive from students. Similarly, evaluators will be able to evaluate systematic trends that persist across semesters.

Concluding Recommendations

At the beginning of this chapter I suggested that a multimethod approach to faculty evaluation should be employed. Therefore, SETs should be used as only one measure of faculty effectiveness and they should be interpreted in context. In fact, Smith (2008) proposed a multidimensional model that includes self reflection, student learning, peer review, and student feedback as potential sources of data that can be used to evaluate teaching effectiveness.

To ensure they are useful indicators of faculty effectiveness, both qualitative and quantitative SETs, should be considered in context. Interpreting only the quantitative or the qualitative data independently can lead to inaccurate interpretations. Similarly discrete analyses that do not include a faculty members self-evaluation or influencing contextual factors may lead to inaccurate appraisals. The developmental progression of the faculty member also must be considered, along with their teaching philosophy, and the situational context. If for example, the faculty member received consistent feedback that improvements were needed across a number of evaluation periods, but this feedback went unheeded, then the ongoing issues must be addressed. Similarly, if the faculty member received feedback, interpreted the feedback, and engaged in revising the course accordingly, then the annual appraisal should reflect that improvement in instruction. In other words, engaging in self-evaluation and modifications, despite some negative SETs, can be evaluated positively. This nuanced approach to evaluation necessarily imposes an even greater onus on the evaluator to remain vigilant in providing constructive feedback and direction.

Halonen, Dunn, McCarthy, and Baker (in press) provide guidance for documenting teaching effectiveness and evaluators may use their recommendations for faculty as a guide for providing feedback to faculty during the evaluation process. They suggest that faculty should be attentive to ongoing trends, departmental norms, individual circumstances, and situational context, and should address negative feedback directly. Indeed, faculty can use these suggestions as an outline for engaging in reflective self-evaluation.

Use SETs to evaluate and improve instruction. All too often faculty and evaluators treat the annual review process as a single event. Smith (2008) offers a comprehensive approach to improving instruction in connection with the evaluation process. His approach is intensive and includes multiple measures. Although this comprehensive model may be daunting, at a minimum, if the process of annual review is to achieve the goal of improving instruction, then specific recommendations for instructional improvement should emerge from the review process. Evaluators can provide guidance for developing one's professional competencies or addressing the affective elements of teaching. Specific strategies can be developed and the effectiveness of the changes can be considered in a subsequent review period.

Use multiple sources of data. Throughout this chapter I suggest that quantitative and qualitative SETs should be interpreted in the context of the institution, department, and situational factors unique to each instructor. It should also be apparent that additional measures of faculty performance and student learning inform the evaluation process (Hoyt & Pallett, n.d.). Evaluators should consider trends in multiple sources of data (e.g., withdrawal rates, teaching philosophy) that are present across courses and semesters. Evaluators should also consider recommendations from prior annual reviews, along with faculty self-assessment measures, in the overall promotion and tenure process.

Develop clear criteria for how SETs will be used in the evaluation process On the one hand; SETs are purportedly used for improving instruction. Ideally, individual faculty members routinely and thoughtfully consider student feedback as one measure of teaching effectiveness that will inform ongoing improvement in instruction. SETs are also used as a measure of teaching effectiveness during the annual review process. If SETs are to be used equitably, then evaluators should establish clear criteria for how they will be interpreted and used in the evaluation process (Cashin, 1996).

Ultimately SETs should be used to improve instruction. Although this chapter addresses the role of SETs in the evaluation process, evaluation necessarily contains an iterative process that involves a mechanism for instructional improvement. Evaluators have a responsibility not only to assess teaching effectiveness as one measure of student learning, but also to provide faculty with development support (Smith, 2008). Good teaching involves specific professional competencies and affective dimensions. Effective administrators should find ways to provide faculty with professional development opportunities that help faculty develop on both dimensions. For example, it may be possible to send faculty to a regional teaching conference or to access professional development opportunities through local campus-based resources (e.g., center for excellence in teaching and learning). Faculty may also wish to institute mid-semester reviews to obtain feedback early in the process.

References

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*, 193-209.

Ballard, M., Rearden, J., & Nelson, L. (1976). Student and peer rating of faculty. *Teaching of Psychology, 3*, 88-90.

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*, 656-665. doi:10.1037//0022-0663.87.4.656

Beyers, C. (2008). The hermeneutics of student evaluations. *College Teaching, 56*, 102-106. doi:10.3200/CTCH.56.2.102-106

Buskist, W. (2002). Effective teaching: Perspectives and insights from Division Two's 2- and 4-year awardees. *Teaching of Psychology, 29*, 188-193. doi:10.1207/S15328023TOP2903_01

Buskist, W., Keeley, J., & Irons, J. (2006). Evaluating and improving your teaching, *Observer, 19*(4), 27-30.

Cashin, W. (1996). *Developing an effective faculty evaluation system* (IDEA Paper #33). Retrieved from http://www.theideacenter.org/sites/default/files/Idea_Paper_33.pdf

Diamond, R. M. (2008). *Designing and assessing courses and curricula: A practical guide* (3rd ed.). San Francisco, CA: Jossey-Bass.

Dunn, D. S., McCarthy, M. A., Baker, S., Halonen, J. S., & Hill, G. W. (2007). Quality benchmarks in undergraduate programs. *American Psychologist, 62,* 650-670.

Dunn, D. S., McCarthy, M. A., Baker, S. C., Halonen, J. S. (2011). *Using quality benchmarks for assessing and developing undergraduate programs*. San Francisco, CA: Jossey-Bass.

Eckert, J. M., & Dabrowski, J. (2010). Should value-added measures be used for performance pay? *Kappan, 91*(8), 88-92.

Germaine, M., & Scandura, T. A. (2005). Grade inflation and student individual differences as systematic bias in faculty evaluations. *Journal of Instructional Psychology, 32*, 58-67.

Greenwald, A. G. (2007). Validity concerns and usefulness of student ratings on instruction. *American Psychologist, 103,* 1182-1186. doi:10.1037//0003-066X.52.11.1182

Halonen, J. S., Dunn, D. S., McCarthy, M. A., & Baker, S. C. (in press). Are you really above average? Documenting your teaching effectiveness. To appear in B.Schwartz & R. A. R. Gurung (Eds.), Evidence-based teaching for higher education. Washington, DC: APA Books.

Hoyt, D. P. & Pallett, W. H. (n.d.). *Appraising teaching effectiveness: Beyond student ratings* (IDEA Paper #36). Retrieved from http://www.theideacenter.org/sites/default/files/Idea_Paper_36.pdf

Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the teacher behavior checklist. *Teaching of Psychology, 37*, 16-20. doi:10.1080/00986280903426282

Keeley, J., Smith, D., & Buskist, W. (2006). The teacher behavior checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33*, 84-91. doi:10.1207/s15328023top3302_1

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal Educational Psychology, 52*, 77-95. doi:10.1111/j.2044-8279.1982.tb02505.x

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal, 21*, 341-366.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253-388.

Marsh, H. W., & Roche, L. A. (1997). Making students evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187-1197. doi:10.1037/0003-066X.52.11.1187

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225. doi:10.1037//0003-066X.52.11.1218

Myers, D. G. (2005, March). Teaching tips from experienced teachers. *Observer, 18*(3). Retrieved from http://www.psychologicalscience.org/observer/getArticle.cfm?id=1745

Price, C. (Summer, 2009). Why don't my students think I'm groovy? The new R's for engaging millennial learners. *Psychology Teacher Network, 19(2)*, 1, 3-5.

Pusateri, T. P. (in press). Contributing psychological expertise to institutional outcomes assessment initiatives. To appear in D. S. Dunn, S. C. Baker, C. M. Mehrotra, R. E. Landrum, & M. A. McCarthy (Eds.), *Assessing teaching and learning in psychology: Current and future perspectives.* Belmont, CA: Cengage.

Saville, B. (2008). *Knowing psychology, teaching psychology: We preach, but do we practice?* Southeastern Conference on the Teaching of Psychology, Atlanta, GA.

Smith, C. (2008). Building effectiveness in teaching through targeted evaluation and response: Connecting evaluation to teaching improvement in higher education. *Assessment & Evaluation in Higher Education, 33*, 517-533. doi:10.1080/02602930701698942

Zinn, T. (2008). *Living on the edge: Embracing risk*. Southeastern Conference on the Teaching of Psychology, Atlanta, GA.

Contact Information

Maureen McCarthy may be contacted at Maureen_McCarthy@kennesaw.edu

Bias in Student Evaluations

Susan A. Basow and Julie L. Martin
Lafayette College

The question of whether student evaluations can be biased is a critical one for those using them, whether for formative or summative purposes. If student evaluations reflect more than an instructor's actual teaching ability, such as how attractive the professor is, this information must be taken into account when such evaluations are used. Given the important role student evaluations play in many academic employment decisions--such as hiring, promotion, tenure, salary, and awards--it is vital to understand potential sources of bias (see McCarthy, this volume). In this chapter, we will examine potential biasing factors involving the professor--such as gender, race/ethnicity, attractiveness, and age--as well as the course, such as course difficulty and expected grade.

Instructor Factors

Social psychologists have documented how a rater's perception of and reaction to another person can be affected by bias, either consciously or unconsciously (Biernat, 2003; Eagly & Karau, 2002; Phelan, Moss-Racusin, & Rudman, 2008.) In particular, cultural stereotypes, such as for gender and race, may create different expectations for different individuals. For example, because women are expected to be nurturant and caring, a woman's interpersonal skills may be viewed more critically in a rater's overall evaluation than would skills of her male counterpart. Furthermore, women and minorities often must work harder to be perceived as equally competent as White men (the normative group), and it is far easier for them to "fall from grace" as well (Biernat, Fuegen, & Kobrynowicz, 2010; Foschi, 2000). Thus, students might perceive the same behavior, such as grading harshly, more negatively if the professor is a woman or African American or Hispanic (who "should" be "nice" and "caring") than if the professor is a White man (who has greater legitimacy due to both race and gender). Such indeed seems to be the case as we discuss below.

Documenting potential sources of bias in field research is challenging because professors vary on many factors, all at the same time. That is, professors not only have a gender, they also have a race, an ethnicity, a certain personality and speaking style; they vary in age and type of course taught in terms of size, level, and discipline. Because it is likely that many of these factors interact to create a particular impression, it is difficult to tease apart only the effects of age or race or gender, etc. In order to do that, laboratory studies often are performed, but their external validity then may be questioned. We will examine both laboratory and field research on this issue in order to explore the possible biasing effects of a professor's gender, race/ethnicity, age, and level of attractiveness on student ratings of teacher effectiveness. We will also attend to the actual questions being asked, because different aspects of teaching effectiveness (e.g., knowledge, dynamism, concern about students) may show different patterns.

Instructor gender

Many research studies have examined whether faculty gender affects student ratings and generally results suggest the negative (e.g., Bennett, 1982; Feldman, 1993); that is, women faculty do not appear to get lower evaluations than do male faculty across the board. This seemingly reassuring result, however, is deceptive because gender appears to operate in interaction with other variables, such as the gender of the rater, the gender-typing of the field in which one teaches, one's gender-typed

characteristics, and status cues. Although gender effects, when found, generally are small in size, they still may have an impact on the ratings received by some women faculty.

The most frequent finding is that teacher gender interacts with student gender to influence student ratings. Whereas male faculty tend to be rated similarly by their male and female students, female faculty tend to be rated lower by their male students and sometimes higher by their female students (Basow, 1995; Basow & Silberg, 1987; Centra & Gaubatz, 2000; Feldman, 1993). The male students who are most likely to devalue their female professors tend to be business and engineering majors, students who tend to hold the most traditional attitudes toward women. Although male students are more likely to rate their female professors lower than their male professors and are less likely to consider them one of their "best" professors, they are not more likely to consider them their "worst" professor (Basow, 2000; Basow, Phelan, & Capotosto, 2006). In contrast, female students often do choose women faculty as "best" and rate them higher than male faculty, especially on qualities related to "fairness" and "providing a comfortable classroom environment." In general, men faculty often are rated higher than women on questions related to scholarship/knowledge and dynamism/enthusiasm, while women faculty often are rated higher than men on questions relating to faculty-student interactions and quality (Bachen, McLoughlin, & Garcia, 1999; Basow & Montgomery, 2005; Bennett, 1982). For example, in Centra and Gaubatz's (2000) study of 741 classes at 21 institutions, male professors were evaluated similarly by their male and female students, but female professors were rated higher by their female students overall and on questions relating to communication and faculty-student interaction.

The subject matter that a professor teaches also plays a role in student ratings. Overall, humanities professors tend to get higher ratings with natural science and engineering professors getting the lowest ratings (Basow, 1995; Basow & Montgomery, 2005). Teacher gender tends to interact with student gender in the humanities and social sciences, with female faculty receiving lower ratings from their male students than they do from their female students. But in the natural sciences, all students tend to rate female faculty lower than male faculty, especially on questions such as "demonstrates knowledge." This result may be due to the fact that the sciences are considered traditionally masculine fields.

Gender as well as discipline may affect a particular professor's teaching style. For example, men are more likely than women to use a lecture-based teaching style, perhaps because they are more likely to be teaching fact-based courses, such as the physical sciences (Basow & Montgomery, 2005; Brady & Eisler, 1999; Canada & Pringle, 1995). Conversely, women are more likely than men to use a more discussion-based teaching style, perhaps because they are more likely to be teaching humanities courses. Still, even when faculty members are matched in terms of rank and discipline, female faculty are found to be more student-oriented and to engage students more in discussions than their male counterparts (Statham, Richardson, & Cook, 1991). In contrast, male faculty appear more likely than female faculty to assert their authority in the classroom through public reprimands and corrections. It may be that these different teaching styles appear gendered to students as well such that women who use a lecture-based teaching style are evaluated more negatively than men who do so.

Teacher personality characteristics also may affect student evaluations in gendered ways. Because women are expected to be caring, they are judged more critically than their male counterparts when they appear to violate students' gendered expectations, such as by being demanding, grading harshly, not accepting student excuses, and not being always available (Basow et al., 2006; Bennett, 1982; Sinclair & Kunda, 2000). Sprague and Massoni (2005) in their qualitative study found that students expect more of their women professors compared to their men professors in terms of time and help and react with greater hostility if these expectations are not met.

In general, women faculty bear the burden of higher expectations. Because gender expectations of men overlap considerably with expectations of professors (e.g., competence, knowledge, high status), a male professor has credibility regardless of age or appearance. For example, Arbuckle and Williams (2003) found that students rated a "young" male professor higher than they rated a "young" female professor in a laboratory study that used the exact same lecture but varied the description of the professor in terms of age and gender. Thus age (and other low status cues) may interact with gender to affect student ratings of women and not men. This may be because gender expectations of women do not overlap much with expectations of professors. Therefore, women professors often have to "prove" that they are credible by more clearly displaying expected "professor" qualities of knowledge, competence, and assertiveness, but must do so while also displaying expected "feminine" qualities, such as warmth and nurturance.

In summary, women faculty are expected to be more available and more nurturing than men faculty, and they typically are. But these qualities only result in comparable evaluations, not higher ones. If, however, female professors are not *more* available and nurturant than their male counterparts, such as by having more office hours or requiring less work, they will be rated *lower* than similar male colleagues. Thus, comparable ratings of male and female faculty may mask a differential set of student expectations for faculty behavior. For particular women (e.g., those who appear young, who are in a stereotypically masculine field, who have a no-nonsense teaching style and teach primarily male students), gender variables can have a negative impact on their student evaluations.

Professor race and ethnicity

Relative to professor gender, the effects of professor race and ethnicity on student evaluations have not been widely studied. This may be due to the relatively low percentage of non-White faculty at institutions of higher education in the U.S. Similar to instructor gender, however, it is likely that racial and ethnic stereotypes affect students' perceptions of and reactions to minority faculty. In particular, African American and Hispanic professors are likely to have to "prove" their knowledge and competence in ways that White professors do not. It also is likely that instructor race/ethnicity interacts with professor gender as well as with student race/ethnicity but very few studies have examined these interactions.

Some research suggests that minority professors may enact their roles differently than White professors. For example, Harlow's (2003) interviews with White and African American faculty members at a large predominantly-White state university found Black professors showed disproportionate amounts of doubt, questioning of their own status, and feeling the need to prove their abilities, as compared to the White professors. Minority faculty, especially women, also appear to have a greater service and mentoring burden than their White counterparts (Griffin & Reddick, 2011) that may translate into higher student expectations of these qualities in minority faculty, similar to those found for White women.

In general, African American and Hispanic faculty appear to receive lower evaluations than White and Asian faculty (Hamermesh & Parker, 2005). For example, using student evaluations of faculty at the top 25 liberal arts colleges in the U.S. posted on the website ratemyprofessor.com, Reid (2010) found that Black faculty, especially Black men, were evaluated more critically and given lower ratings on quality, helpfulness, and clarity than their White counterparts. Similarly, Smith (2007) found that White faculty were rated consistently higher than Black faculty on global measures of overall teaching at a large university in the southern United States. Because these are naturalistic studies, it is not clear whether such differential ratings are due to bias, to actual differences in teaching effectiveness, or to other

factors, such as the subject matter being taught or teaching style. Some indication that bias may be involved comes from a study by Ho, Thomsen, and Sidanius (2009) who examined how ratings of professor intellectual competence and sensitivity to students related to ratings of overall teaching effectiveness in a sample of 5,655 randomly-selected students. Although the overall performance ratings of African American faculty did not differ from those of White faculty, students' perceptions of intellectual competence were a bigger factor in the overall performance evaluations of Black compared to White faculty. This finding is consistent with social psychological research findings that lower status groups (e.g., women, African Americans) must "prove" competence when evaluated for high status positions (e.g., Foschi, 2000).

In one of the few laboratory studies on the effects of professor race (White, Asian, or African American) and gender on student evaluations, Bavishi, Madera, and Hebl (2010) asked entering college students to rate a Curricula Vita as to the competency, legitimacy, and interpersonal skills of the hypothetical professor. Results revealed that African American professors, especially women, were rated the lowest on all three dimensions and Asian professors were rated lower than the White professors on interpersonal skills. Therefore, although Asian professors may be viewed more positively than Black professors, they still may experience some negative perceptions based on race.

The type of course that minority faculty members teach may also affect how students perceive and rate them. Given that White males are viewed as the "normative" professor, both women and minority faculty may be viewed both as less legitimate and less objective. In race-focused diversity courses, most likely to be taught by minority faculty, students appear to view African American faculty as more biased and subjective, although more knowledgeable, than White faculty teaching the same course (Anderson & Smith, 2005; Littleford, Ong, Tseng, Milliken, & Humy, 2010). In general, faculty teaching about White privilege to White students often receive lower student evaluations in those courses than in their other courses (Boatright-Horowitz & Soeung, 2009), a finding that may contribute to the lower ratings of African American and Hispanic faculty frequently found.

Given the paucity of minorities among the professorate as well as the student body, it is not clear whether minority students react the same way to minority professors, especially of their own race/ethnicity, as White students do. In a few naturalistic studies, students have been found to receive more attention and support by a faculty member of the same race than by a faculty member of another race (Dee, 2005; Ehrenberg, Goldhaber, & Brewer, 1995). It is possible that teachers' in-group support may influence students' perceptions and evaluations such that a student in-group preference is found, at least on some questions. As noted above, this frequently is the pattern found regarding how student gender and faculty gender interact. In other contexts, such as doctor-patient relationships and peer friendships, race-concordance is associated with more positive ratings (Cooper-Patrick et al., 1999; Verkuyten, 2007). Research is needed on this question.

Overall, it is likely that race/ethnicity of the professor affects student ratings, with White faculty generally rated higher than minority faculty, but this effect may depend on the specific question (e.g., overall, interpersonal) and other variables, such a professor gender, student race/ethnicity, and type of course being evaluated.

Professor age

Another demographic variable that has been insufficiently studied is professor age. This may be particularly salient as the baby-boom generation, now entering retirement, moves through the

professorate. As with other professor factors, it is likely to operate in interaction with other variables, such as professor gender, student age, and type of course taught.

Age discrimination has been found in the workplace, with older job applicants often viewed more negatively than younger and middle-aged applicants (e.g., Kite, Stockdale, Whitley, & Johnson, 2005; Krings, Sczesny, & Kluge, 2011). In order to examine the effect of professor age on student evaluations, several laboratory studies have been conducted, wherein a professor's age is varied by either pictures and/or written descriptions. In two studies using a hypothetical male professor of three different ages (25, 53, 73), college students tended to rate the oldest professor most negatively (Levin, 1988; Stolte, 1996). In Arbuckle and Williams' (2003) experimental study, female and male students watched slides of an age- and gender-neutral stick figure and listened to a neutral voice presenting a lecture. Students then evaluated the stick-figure professor on teacher evaluation forms that indicated one of four different age and gender conditions—male or female; "old" (over 55) or "young" (under 35). The "young" male professor received significantly higher ratings on questions tapping "enthusiasm" and "meaningful voice tone" than did the "young" female professor and both "old" professors. There was no effect of age or gender on ratings of "seemed to be relaxed and confident." Thus age and gender may interact to the benefit of younger males, at least on some evaluative questions. Because these are all laboratory studies and two of the three are over 15 years old, it is not clear how well they relate to contemporary evaluation contexts.

In the only quasi-naturalistic study found that explored the effect of professor age, Radmacher and Martin (2001) recruited a volunteer sample of 13 professors (2 male, 11 female) from different disciplines to collect mid-term evaluations from their students (N = 351). Results indicated that teacher age was significantly negatively correlated with ratings of teacher effectiveness, but the size of the correlation was small. Given the double-standard of aging in U.S. culture wherein older women are viewed even more negatively than older men (Kite et al., 2005), and the predominance of women professors in the Radmacher and Martin study, it is possible that instructor age and gender may interact.

Overall, professor age is understudied and likely to operate in interaction with other factors in affecting student evaluations. There is some evidence that older faculty may receive "lower" evaluations relative to "younger" ones. Whether this relationship is curvilinear (i.e., perhaps the best-rated professor is in her/his 40s rather than 30s or 50s) needs to be examined further.

Professor attractiveness

Professor attractiveness is a more seemingly-subjective variable than the other categories of potential bias (gender, race/ethnicity, age) we have discussed. Considerable research has demonstrated that attractive people generally are liked better and perceived more positively than their less-attractive peers (Dion, Berscheid, & Walster, 1972; Langlois, Kalakanis, Rubenstein, Larson, Hallam, & Smoot, 2000). This preferential treatment of attractive people holds true for both females and males and has been found in student evaluations as well.

Economists Hamermesh and Parker (2005) took pictures of professors from departmental websites and various other sources at the University of Texas in Austin and had six students (3 male, 3 female) rate the attractiveness of each picture. Using each professor's average attractiveness rating along with other demographic and course variables, the researchers examined which factors significantly predicted the end-of-term ratings of course excellence. Results indicated that the effects of professor attractiveness on average course ratings were significant and strong, especially in lower division courses. Moving from

one standard deviation below the mean in attractiveness to one standard deviation above led to nearly a full standard deviation increase in the average class rating of teacher effectiveness. Interestingly, attractiveness ratings accounted for more of the variance in the ratings of male compared to female professors, perhaps because women were rated lower than men on teaching effectiveness, or because other factors may contribute more to ratings of women than to men (e.g., personal traits, teaching style, type of course, gender of student).

Other research has found similar results. For example, using student ratings from the website ratemyprofessor.com, Riniolo and colleagues (2006) found that those teachers with higher scores on the attractiveness scale also had higher scores on ratings of teacher effectiveness. This significant positive correlation was found for both men and women. Although data on this website is not a random sample, the findings are consistent with other research. Overall, physical attractiveness of a professor may bias student evaluations in a positive direction. Further research is needed to examine the possible interaction between professor gender and professor attractiveness on students' evaluations.

Course-Related Factors

As student evaluations have increased in importance in employment-related decisions, so has concern over whether faculty can improve the ratings they receive by reducing the rigor of the course material or by grading more leniently. These two factors have received extensive study and the answer is complicated (see also Keeley, this volume and Wilson & Ryan, this volume).

Expected grade

Some relationship between grades students receive and student evaluations of teacher effectiveness is to be expected. Indeed, one way to assess the validity of student evaluations is to ascertain if the teachers with the highest ratings also have students who learn more, as assessed by their academic performance and the grades they receive. The best test of this relationship is in multiple-section courses taught by different instructors where all students take a common final exam. In such studies, there is a modest but significant positive correlation between student academic performance and student evaluations (Cohen, 1981).

In other research designs, however, where there is no common course material or common exam, the relationship between student grade and teacher evaluations may reflect something other than teaching effectiveness—it may reflect student "liking," or gratefulness for a high grade, or student unhappiness and dislike for a low grade, regardless of actual student learning. Especially for untenured faculty, this presumed relationship has been viewed as contributing to grade inflation (Eizler, 2002)—an increase in grading leniency in order to obtain more positive student evaluations.

This assumed grading effect is a source of great concern for those who have to interpret and utilize student evaluations. In order to assess the research on whether such a grading effect occurs, it is important to make a distinction between actual and expected grade. Since course evaluations typically are completed before students receive their final course grade, a grading effect would be demonstrated by a greater correlation between students' expected grade and student ratings of their professor than between students' actual grade and their ratings. Such indeed is the case (Felton, 2008; Greenwald & Gillmore, 1997; Millea & Grimes, 2002). For example, Ducette and Kenney's (1982) examination of 456 classes at an eastern university over the course of three years found that students who expected higher grades gave their instructor higher ratings on effectiveness than those who expected lower grades. However, there was no significant relationship between actual grades and ratings of teacher effectiveness.

Eizler (2002) investigated whether the use of student evaluations of teaching effectiveness contributed to grade inflation by examining student evaluations in more than 37,000 course sections between 1980 and 1999 in a mid-sized, public university in the upper Midwest. The percentage of students expecting A/A– grades was relatively stable during the 1980s but increased steadily by more than 10% over the next 10 years; the same pattern was found in student ratings of teaching (i.e., student ratings were relatively stable during the 1980s but increased steadily in the '90s.) The correlations between expected grade and student ratings remained significant even after controlling for variables tapping alternative explanations, such as prior achievement, course popularity and instructor appeal. Thus, it seems likely that grading leniency can bias student ratings in the positive direction.

Course difficulty

Course difficulty is another course factor that may affect student ratings either directly (if students give higher ratings to professors of "easier" courses than to those of more challenging ones) or indirectly through its effect on grades (i.e., students in "easier" courses may expect higher grades, and it is this higher expected grade that is associated with higher student ratings). The research on this question does not present a clear picture.

Some studies find a direct relationship (e.g., Addison, Best, & Warrington, 2006) wherein courses considered easier than expected receive higher ratings than courses viewed as harder than expected, regardless of student grade. Other studies find evidence of an indirect relationship between ratings of course difficulty and student evaluations, mediated by expected course grade (e.g., Ducette & Kenney, 1982). Still other studies find either no relationship or a more complicated one (Heckert, Latier, Ringwald-Burton, & Drazen, 2006; Marsh & Roche, 2000; Millea & Grimes, 2002; Zabaleta, 2000). For example, there may be a relationship between perceived course difficulty and student ratings only for students expecting lower grades. In some cases, greater student effort is associated with higher rather than lower student ratings.

Summary and Implications

Although there is still a considerable amount of research needed to understand all the ways that student evaluations can be biased, this chapter suggests that not only is some bias possible but it is likely. As a human activity reliant upon person perception and interpersonal judgment, student ratings are affected by the same factors that can potentially affect any rater's judgment: stereotypes based on gender, race/ethnicity, age, and other qualities (such as professor sexual orientation); the equation of "what is beautiful is good;" more positive feelings towards those who seem to reward us (e.g., with good grades). Even though the size of individual effects may be small, for specific professors these small effects may add up to make a meaningful difference on the ratings they receive. Although the average-looking young-to-middle-aged White male professor teaching traditional courses may receive student ratings that are relatively unbiased reflections of his teaching effectiveness, other professors (women, minorities, older, unattractive-looking, teaching diversity-related courses) may receive evaluations that reflect some degree of bias. It behooves those who use such ratings for evaluative purposes to understand the subtle ways such variables may operate, especially in interaction with each other.

References

Addison, W.E., Best, & Warrington, J.D. (2006). Students' perceptions of course difficulty and their ratings of the instructor. *College Student Journal, 40*, 409-416.

Anderson, K. L., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences, 27,* 184-201. 10.1177/0739986304273707

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles, 49*, 507-516. 10.1023/A:1025832707002

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*, 193-210. 10.1080/03634529909379169

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*, 656-665. 10.1037/0022-0663.87.4.656

Basow, S. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles, 43*, 139-149. 10.1023/A:1026655528055

Basow, S. A., & Montgomery, S. (2005). Student evaluations of professors and professor self-ratings: Gender and divisional patterns. *Journal of Personnel Evaluation in Education, 18*, 91-106. 10.1007/s11092-006-9001-8

Basow, S. A., Phelan, J., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30*, 25-35. 10.1111/j.1471-6402.2006.00259.x

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*, 308-314. 10.1037/0022-0663.79.3.308

Bavishi, A., Madera, J. M., & Hebl, M. R (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education. 3*, 245-256. 10.1037/a0020763

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluations. *Journal of Educational Psychology, 74,* 170-179. 10.1037/0022-0663.74.2.170

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58*, 1019-1027. 10.1037/0003-066X.58.12.1019

Biernat, M., Fuegen, K., & Kobrynowicz, D. (2010). Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. *Personality and Social Psychology Bulletin, 36*, 855-868. 10.1177/0146167210369483

Boatright-Horowitz, S. L., & Soeung, S. (2009). Teaching White privilege to White students can mean saying good-bye to positive student evaluations. *American Psychologist*, *64*, 574-575. 10.1037/a0016593

Brady, K. L., & Eisler, R. M. (1999). Sex and gender in the college classroom: A quantitative analysis of faculty-student interactions and perceptions. *Journal of Educational Psychology, 91*, 127-145. 10.1037/0022-0663.91.1.127

Canada, K., & Pringle, R. (1995). The role of gender in college classroom interactions: A social context approach. *Sociology of Education, 68*, 161-186. 10.2307/2112683

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education, 71*, 17-33. 10.2307/2649280

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281-309.

Cooper-Patrick, L., Gallo, J. J., Gonzales, J. J., Vu, H. T., Powe, N. R., & Nelson, C., etal. (1999). Race, gender, and partnership in the patient-physician relationship. *Journal of the American Medical Association, 282*, 583-589. 10.1001/JAMA.282.6.583

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review, 95*, 158-165. 10.1257/000282805774670446

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 2,* 285-290. 10.1037/h0033731

DuCette, J., & Kenney, J. (1982). Do grading standards affect student evaluations of teaching? Some new evidence on an old question. *Journal of Educational Psychology, 74*, 308-314. 10.1037/0022-0663.74.3.308

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Bulletin, 108*, 233-256. 10.1037/0033-295X.109.3.573

Ehrenberg, R. G., Goldhaber, D.D., & Brewer, D.J. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from the national educational longitudinal study of 1988. *Industrial and Labor Relations Review, 48*, 547-561. http://www.nber.org/papers/w4669

Eizler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43,* 483-501. 10.1023/A:1015579817194

Feldman, K. (1993). College students' views of male and female college teachers: Part II--Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34,* 151-211. 10.1007/BF00992161

Felton, J. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on ratemyprofessors.com. *Assessment and Evaluation in Higher Education, 33*, 45 - 61. 10.1080/02602930601122803

Foschi, M. (2000). Double standards for competence: Theory and research. *Annual Review of Sociology, 26*, 21-42. 10.1146/annurev.soc.26.1.21

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209-1217. 10.1037/0003-066X.52.11.1209

Griffin, K. A., & Reddick, R. J. (2011). Surveillance and sacrifice: Gender differences in the mentoring patterns of Black professors at predominantly White research universities. *American Educational Research Journal, 48*, 1032–1057. 10.3102/0002831211405025

Hamermesh, D. S., & Parker, A. M. (2005). Beauty in the classroom: Professors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*, 369-376. 10.1016/j.econedurev.2004.07.013

Harlow, R. (2003). "Race doesn't matter, but...": The effect of race on professors' experiences and emotion management in the undergraduate college classroom. *Social Psychology Quarterly, 66*, 348-363. 10.2307/1519834

Heckert, T. M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to "buy" better evaluations through lenient grading? *College Student Journal, 40*, 588-596.

Ho, A. K., Thomsen, L., & Sidanius, J. (2009). Perceived academic competence and overall job evaluations: Students' evaluations of African American and European American professors. *Journal of Applied Social Psychology, 39*, 389–406. 10.1111/j.1559-1816.2008.00443.x

Kite, M. E., Stockdale, G. D., Whitley, B. E., & Johnson, B. T. (2005). Attitudes toward younger and older adults: An updated meta-analytic review. *Journal of Social Issues, 61*, 241-266. 10.1111/j.1540-4560.2005.00404.x

Krings, F., Sczesny, S., & Kluge, A. (2011). Stereotypical inferences as mediators of age discrimination: The role of competence and warmth. *British Journal of Management. 22,* 187-201. 10.1111/j.1467-8551.2010.00721.x

Langlois, J. H., Kalakanis, L., Rubenstein, A.J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126,* 390-423. 10.1037/0033-2909.126.3.390

Levin, W. C. (1988). Age stereotyping: College student evaluations. *Research on Aging, 10*, 134-148. 10.1177/0164027588101007

Littleford, L. N., Ong, K. S., Tseng, A., Milliken, J. C., & Humy, S. L. (2010). Perceptions of European American and African American instructors teaching race-focused courses. *Journal of Diversity in Higher Education, 3*, 230–244. 10.1037/a0020950

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92*, 202-228. 10.1037/0022-0663.92.1.202

Millea, M., & Grimes, P. W. (2002). Grade expectations and student evaluation of teaching. *College Student Journal, 36*, 582-590. 0146-3934

Phelan, J. E., Moss-Racusin, C. A., & Rudman, L. A. (2008). Competent yet out in the cold: Shifting criteria for hiring reflect backlash toward agentic women. *Psychology of Women Quarterly, 32*, 406-413. 10.1111/j.1471-6402.2008.00454.x

Radmacher, S. A., & Martin, D. J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *Journal of Psychology, 135*, 259-268. 10.1080/00223980109603696

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education, 3*, 137-153. 10.1037/a0019865

Riniolo, T. D., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *Journal of General Psychology, 133*, 19–35. 16475667

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality & Social Psychology Bulletin, 26*, 1329-1342. 10.1177/0146167200263002

Smith, B. P. (2007). Student ratings of teaching effectiveness: An analysis of end of course faculty evaluations. *College Student Journal, 41*, 788-800. 0146-3934

Sprague, J., & Massoni, K. (2005). Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles, 53*, 779-793. 10.1007/s11199-005-8292-4

Statham, A., Richardson, L., & Cook, J. (1991). *Gender and university teaching: A negotiated difference.* Albany, NY: State University of New York Press.

Stolte, J. F. (1996). Evaluation of persons of varying ages. *Journal of Social Psychology. 136*, 305-309. 10.1080/00224545.1996.9714009

Verkuyten, M. (2007). Ethnic in-group favoritism among minority and majority groups: Testing the self-esteem hypothesis among preadolescents. *Journal of Applied Social Psychology, 37*, 486-500. 10.1111/j.1559-1816.2007.00170.x

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education, 12*, 55-76. 10.1080/13562510601102131

Contact Information

Susan Basow may be contacted at basows@lafayette.edu

On-line Measures of Student Evaluation of Instruction

Cheryll M. Adams
Ball State University

Evaluation is a controversial topic in any field when one person is given the charge to evaluate another, particularly when the evaluator is junior in experience to the person being evaluated (Haefele, 1993; Kulik, 2001). The field of education, whether we are targeting K-12 or higher education, has individuals on both sides of the evaluation coin. Some argue for the importance of having students evaluate their instructors while others argue just as strongly that students, who generally do not understand teaching pedagogy, are not competent to determine whether an instructor performs at a particular level. In higher education, evaluations may take a number of forms, both formative and summative (see Addison and Stowell, this volume, and Keeley, this volume). Some instructors use "exit cards" or "tickets to leave" periodically during the semester. That is, instructors may hand out actual pieces of paper with questions about how the student perceives the class thus far or may dictate questions while students answer on index cards or slips of paper. Some instructors may ask questions related to a particular assignment. These instruments are generally used to gather data on how the students feel about the course at a particular time during the semester. These assessments may be collected as often as every few weeks, particularly if they are based on simple queries such as "here's what's working for me," or "here is where I am having trouble." Such assessments allow the instructor to make changes to the course based on feedback before the end of the course. End of course evaluations gather summative data about individual classes and may be used by an instructor to make changes in the course the next time it is taught. Traditionally, these summative evaluations have been used to compare faculty and may be used to make salary, promotion, and tenure decisions. Another controversy is sparked when different departments use different questions on their respective instruments, making a true comparison nearly impossible (see McCarthy, this volume, for a discussion).

When the results of student evaluations are used for promotion, tenure, and salary decisions, faculty often list sampling bias as a factor that might affect average ratings. That is, faculty may believe that only those who are truly fond of an instructor and check all categories with the highest rating or totally dislike an instructor and conversely check the lowest rating for every question will respond to the evaluations, thus skewing the results (Smith, 2011). Another controversy exists when the evaluation includes questions that assess students' evaluation of the course. Students who enjoyed the course may give a high rating while those who disliked it, particularly if it is a required course, may indicate their displeasure by rating the course poorly. Faculty continue to discuss the validity of evaluations that depend on the capriciousness of the students responding to the questions.

In recent years, more institutions of higher education (IHE) have moved from paper and pencil surveys to online evaluations of instruction (Avery, Bryant, Mathios, Kang, & Bell, 2006). This practice has not eliminated the previous controversies mentioned, but instead, it has brought new ones to the forefront. The advantages of using online evaluations include cost-effectiveness, more time for responding, and faster feedback to faculty. The trade off, in general, is a lower response rate for the evaluations (Anderson, Cain, & Bird, 2005; Dommeyer, Baum, Hanna, & Chapman,2004).

This chapter will address survey research and online survey research in general to lay the foundation for a discussion of online measures of instruction. I look at the pros and cons of using online measures of instruction instead of traditional paper and pencil measures. I examine the issue of response rate and offer some recommendations for using online measures of instruction effectively.

Online Survey Research

Survey research in general

Surveys are designed to collect information from a group of individuals (Fowler, 2009). A vital component of any survey is question design. After questions have been developed, a pilot test of the questions using individuals from the target population is a necessary step to ensure that the questions are being understood as the developers intended and that the answers are meaningful. Because those who choose to respond to a survey may differ from the actual population, error can interfere with the accuracy of any inferences made using the data (Fowler, 2009). For example, if the population is all undergraduates at the university, there will be sampling error if no seniors respond. If a survey is designed to gather information about faculty members' teaching ability, bias could be a factor if students' ratings are based on how well they like the faculty members rather than on the quality of their teaching (see Wilson & Ryan, this volume). Another important issue is response rate which, in the context of survey research, refers to the proportion of the number in the sample who complete the survey (Fowler, 2009). Response rates for surveys vary considerably, but those that are administered on-site generally have higher response rates than those that are mailed or completed online (Sue & Ritter, 2007). Hence, faculty and administrators are understandably concerned about whether these lower rates adequately represent the population of students in any given course.

Online survey research

There have been many studies focusing on online surveys. An often-cited reference for looking at electronic surveys in general is the meta-analysis of response rates in electronic surveys which used a wide range of databases to garner 68 surveys from 49 studies that met the researchers' criteria (Cook, Heath, & Thompson, 2000). The results from this comprehensive analysis indicate that the number of contacts such as follow-up letters ($r_{s=}$.560), receiving a personalized letter ($r_{s=}$.524), and receiving prenotification ($r_{s=}$.328) were important if the goal is increased response rates. Those factors that were less relevant were the length of the survey and whether a password was required. Sheehan and McMillan (1999) pointed out the lower response rates for electronic surveys compared to surveys distributed through the mail and urged researchers to begin to determine ways to increase the rate. The results of this meta-analysis (Cook, et al., 2000) provided information that can be useful in raising response rates to electronic surveys.

As with any survey not given in a face-to-face situation, we can never be completely sure that the person for whom the survey is intended is the actual person who is completing the survey. Although there have been studies that provide recommendations to diminish the opportunity to cheat when taking an examination online, to date there have been no studies that looked specifically at this issue in online measures of instruction. If evaluations are not completed onsite so that faculty can be sure the correct students are completing the evaluations, there is no way to prove one way or the other who responded to the online evaluation form. This is one area that still needs examining before we can be sure of the integrity of the results.

Using Online Surveys to Measure Instruction

Measures of instruction

Virtually every institution of higher education uses student completed end-of-course evaluations as part of the documentation necessary for salary, promotion, and tenure decisions. These measures of instruction are almost exclusively in the form of surveys, and the advantages and disadvantages of survey research apply to these evaluations. Although faculty may argue the fact, various studies indicate

that student evaluations are considered to be highly reliable and moderately valid (Aleamoni, 1999; Centra, 1993; Hobson & Talbot, 2001; see also Keeley, this volume). Other evaluations (e.g., peer evaluations, syllabus reviews, and chair evaluations) are, in general, a much less reliable and valid (Centra, 1993; but see Ismail, Buskist, & Groccia, this volume and Schafer, Yost Hammer, & Berntsen, this volume, for a discussion of effective use of these evaluation methods). The effect of students' grades on how they rate the instructor appears to be small or non-existent (V. Johnson, 2002; Gigliotti & Buchtel, 1990; Greenwald & Gillmore, 1997). Although many faculty believe students who do poorly in class will rate their instructors low, there is virtually no difference between the instructor's rating from those who do poorly in class and those who do very well (Liegle & McDonald, 2004). What does seem to affect ratings is how much students perceive they have learned despite their final grade, whether the class was "a breeze" or difficult, and whether the class stimulated them intellectually (Centra, 2003; Remedios & Lieberman, 2008). (See Basow & Martin, this volume, for a discussion of other possible biases affecting student evaluations of faculty.)

Despite the widespread use of these evaluations, students seem to have little knowledge about how the evaluations are used and the role they play in salary, promotion, and tenure decisions (Spencer & Schmelkin, 2002; Galliard, Mitchell, & Kavota, 2006). They often believe that no one except the instructor sees the evaluations and that the instructor is not under any obligation to make changes to the course based on the evaluations (Beran & Rokosh, 2009; Nasser & Fresko, 2002; Spencer & Schmelkin, 2002). In fact, there is some truth to these students' beliefs as research indicates some instructors put little value in evaluations and rarely use the results to revise their courses (Beran & Rokosk, 2009; Nasser & Fresko, 2002; Spencer & Schmelkin, 2002). There is evidence that students have a better chance of completing course evaluations if they believe there is some value in doing so. Several studies suggest that if students perceive their opinions count and understand exactly how the evaluations are used, they will be more likely to complete the evaluation (Spencer & Schmelkin, 2002; Galliard et al., 2006).

Paper and pencil measures of instruction

Until recently, most surveys designed to measure students' perceptions of their courses and instructors were administered via paper and pencil. For example, students might blacken in bubbles on an end-of-course survey using a 5-point Likert scale with anchor points of *strongly agree* (5) and *strongly disagree* (1) to indicate their opinions in response to various questions designed to measure their satisfaction with the course and their instructor. Quite often the students received the survey during class the week before the final exam. The surveys were anonymous and generally had a good response rate due to the "captive audience." Although a student could simply not respond by placing a blank survey in the envelope, response rates were usually upwards of 70% (Dommeyer, et al., 2004; Layne, DeCristoforo, & McGinty, 1999). Even with the high response rates of surveying students on-site, the cost of preparing the actual forms, the time necessary for administrative assistants to sort the forms according to instructor and class, the chance of instructor influence on the results, the use of class time to administer the survey, and the length of time necessary to type student responses so instructors cannot recognize handwriting, tabulate and analyze the forms are only a few of the disadvantages in using this method (Dommeyer, et al., 2004; Gamliel & Davidovitz, 2005; Nulty, 2008). Thus advances in technology led many IHE to seriously consider using online course and instructor evaluations.

Online measures of instruction

Online evaluations are those that are accessible to students through a link generally sent to them in an email that includes directions about how to complete the evaluation. A student may receive a notice for each class in which the student is enrolled or all classes may be accessed through a link specifically for

that student. The emails may be sent as a blast to all students simultaneously or directions may be provided at a central location easily accessible to students. Students respond to each question by selecting the appropriate answer and most surveys have a section for open-ended comments. The means for sending the emails to the students and for receiving and analyzing the data are specific to each university or college. Unless the IHE chooses to revise their surveys (other than to change format), the same survey used for the paper and pencil mode can be used for the online mode of evaluation.

The move towards online evaluation in the last decade is evident at IHE, and the body of research on the topic of online evaluations is expanding. From a review of the literature, two major advantages of online evaluations that distinguish them from paper and pencil modes are consistently mentioned: efficiency and cost-effectiveness (Anderson, et al., 2005; Cook, et al., 2000; Dommeyer, et al., 2004; Layne, et.al, 1999; Nulty, 2008; "Student evaluations of teaching," 2011).

The efficiency and cost effectiveness of online evaluations are well-documented (Anderson, et al., 2005; Cook, et al., 2000; Dommeyer, et al., 2004; Layne, et al., 1999; Nulty, 2008; "Student evaluations of teaching," 2011). Although paper and pencil evaluations take an inordinate amount of time to prepare and distribute, online evaluations do not involve hard copies and distribution is at the touch of a button. No class time is wasted by the need to complete paper and pencil copies and the need for someone to monitor the process during class is eliminated. Even with paper evaluations that are scanned rather than being processed by hand, time is a considerable factor leading to inefficiency. Moreover, online evaluations allow for faster analysis and reporting of data. Instead of taking weeks to receive a tabulated report, faculty can receive the information shortly after the evaluation window closes. Many IHE do, however, have a policy concerning when evaluation data can be distributed (e.g., after grades have been submitted) rather than allowing results to be accessible immediately. Online evaluations also are less open to instructor influence. Another advantage is that all students have the opportunity to provide the evaluation when it is convenient for them during the window given for completion. In contrast, when in class evaluations are collected, students who are absent during the day do not have the opportunity to provide an evaluation. Another issue is that the laborious task of getting paper evaluations ready is inefficient and costly. Personnel may need to be pulled from other areas to assist, abandoning their usual tasks. Sometimes additional personnel must be hired to assist for the preparation period. Storage space at most IHE is at a premium and only limited extra space is available in which to store paper surveys until the limit for keeping them expires. The cost is much lower for managing, analyzing, and reporting data that can be stored electronically.

A third advantage reported by several studies looks at benefits to the students. Students are more likely to complete the comment sections because their handwriting cannot be recognized, and they generally view typing comments as an easier task than writing responses (Bullock, 2003; Dommeyer, Baum, Chapman, & Hanna, 2002; Layne, et al., 1999).

Response Rates: Issues and Solutions

Despite the advantages and the fact that the number of IHE that have revised their measures of instruction from paper and pencil to online versions is rapidly increasing, there are still some hurdles to overcome, the most conspicuous of which is the lower response rate for online versions (Anderson, et al., 2005; Avery, et al., 2006; Dommeyer, et al., 2000; Heath, Lawyer, & Rasmussen, 2007; Nulty 2008). Studies show that even when the paper and pencil documents and the online versions were identical, response rates to the online version were lower (Anderson, et al., 2005; "Student evaluations of teaching," 2011). Although Nulty (2008) speculates the issue resides in the face-to-face administration of paper and pencil surveys which, of course, negates one of the main advantages of online surveys, to

date, no one has uncovered the definitive reason or reasons for the lower rate. He does present a table that describes the number of students who must respond for a given class size to procure a specific response rate at a particular confidence level. While Nulty demonstrates his findings for class sizes from 10 to 2000, Table 1 only shows selected class sizes. This table is presented only as a guide, and we are cautioned that both sample error and sample bias will result when response rates are low.

Table 1

Required response rate by class size

	Liberal conditions		Stringent conditions	
	10% Sampling errors; 80% Confidence level		3% Sampling errors; 95% Confidence level	
Total number of students in the course	Required number of respondents	Response rate required	Required number of respondents	Response rate required
10	7	75%	10	100%
20	12	58%	19	97%
30	14	48%	29	96%
50	17	35%	47	95%
70	19	28%	64	91%
100	21	21%	87	87%

Bias

As discussed previously, lower response rates may have both sample and error bias. In a (*N*=2011) study using sex, ethnicity, cumulative GPA, course grade, total number of credits completed at the school, and course size, Jones (2009) found that only course grade was a potential source of sample bias. If only those students who received high course grades responded, the evaluation results may not reflect what the class as a whole perceived about the instructor. Thorpe (2002) also investigated potential sources of bias, and in that study (*N*=844), females, individuals with higher GPAs, and students earning higher grades tended to respond to the request to evaluate courses in greater numbers than did other individuals. Conversely, other researchers have examined the issue of bias and concluded that there is no concrete evidence that online evaluations produce any more bias than paper and pencil evaluations (Avery, et al., 2006; Dommeyer, 2004; Liu, 2006; Sax, Gilmartin, & Bryant, 2003). Of particular note are those studies that used identical surveys given in both formats (Anderson, et al., 2005; "Student evaluations of teaching," 2011). Other than the lower response rates, there were no other significant differences in the responses.

Incentives/Disincentives

Multiple studies suggest the use of incentives (rewards) and disincentives (punitive consequences) as a means of raising response rates (Anderson et al., 2005; Ballantyne, 2003; Dommeyer, et al., 2004; T.

Johnson, 2003; Prunty, 2011; Ravenscroft & Enyeart, 2009). Table 2 provides a list of possible incentives and disincentives with the supporting research. All incentives/disincentives contributed to raising response rates, although some were more effective than others. Contrary to what some faculty may believe, no bias was found in the evaluations when an incentive was used (Dommeyer et al., 2004).

Prunty (2011) describes her own attempt to raise the response rate in her class by offering a 1% overall grade increase if the class response rate reached at least 80%. As with many of the online evaluation systems, the response rate is visible to the faculty member, but the names of the respondents are not. Her students took responsibility to urge each other to complete the evaluations, even bringing in their laptops so fellow students without internet access could complete them. She reported that students took the evaluations more seriously and that they perceived she would take their comments seriously. Consequently, her self-reported response rates range from 80% to 100% each time she implemented the extra credit incentive.

Table 2. Suggested Incentives and Disincentives to Raise Response Rates

Incentive/Disincentive	Supporting Research
Extra credit	Dommeyer, et al., 2004[a] ; Johnson, T.,2003; Prunty, 2011[b]
Early access to course grade	Anderson, Brown, & Spaeth, 2006; Dommeyer, et al., 2004; Johnson, T., 2003.
Course evaluations a part of student grade in course	Ravenscroft & Enyeart, 2009
Donor—either from the community or a university alumnus—contributes one dollar to a local or national charity for every student course evaluation submitted	Ravenscroft & Enyeart, 2009
Prizes: food coupons, university bookstore coupons	Ballantyne, 2003; Johnson,T., 2003

[a] quarter of a percentage point on final grade; [b] 1% grade increase.

Notifications and other factors

Several studies have reported increases in response rates when even small steps such as reminding students about the evaluations are taken. In the analysis completed by Cook, et al. (2000), three factors were instrumental in making greater gains in response rate: the number of contacts made, personalized contacts, and precontacts. Thus frequently reminding students to complete the evaluations,

personalizing the contact rather than just sending a mass mailing, and giving students a "heads up" prior to the beginning of the evaluation window increased student response rates. Additionally, an in-class demonstration of the evaluation system and how it works, thus demystifying the process, had a positive effect on response rates (Dommeyer, et al., 2004). Moreover, Norris and Conn (2005) found that helping students understand the value of the evaluation process also increased response rates. Even these small, virtually no cost measures raised response rates 30 percentage points or more over courses that employed no strategy (Norris & Conn, 2005).

Anonymity as perceived by students is another factor that impacts response rate because they may fear reprisals from faculty if they give them a poor review. When students login with their student identification number, they may feel vulnerable and not believe assurances that neither the university nor the faculty member is tracking their own personal responses (Layne, et. al., 1999). On the bright side, there is evidence that as online evaluations become the norm on campuses and the newness wears off, response rates rise on their own (Avery, et. al, 2006; T. Johnson, 2003).

Recommendations

In summary, most IHE are moving to online end of course evaluations. While the body of research on this topic is rapidly increasing, a change in beliefs held by teachers and students toward online evaluation of teaching is not always moving at a comparable rate. The advantages of cost effectiveness and efficiency of providing evaluations online compared to paper and pencil may not outweigh the drastic drop in response rate usually associated with the former. Because even small means to improve response rates seem to cause them to rise (Norris & Cook, 2005), these ten recommendations are offered as possibilities for implementation.

1. Demonstrate the evaluation system in class.
2. Make students aware of why evaluations are collected, their value, and how they are used.
3. Reassure students about the anonymity and confidentiality of the evaluation results.
4. Send students a pre-evaluation reminder before the evaluation window opens to prepare them for the actual evaluation. (This will also serve as a check to be sure e-mail addresses work and mailboxes are not full.)
5. Make sure students understand how to access the evaluations.
6. Once the evaluation window opens, send students frequent reminders to complete the evaluations; be sure these reminders are only sent to those who have not completed the evaluations, not all students.
7. Encourage faculty members to remind students in class frequently and have those who teach online use the Announcement function in their online platform to persuade students to complete the evaluations.
8. Encourage individual faculty members to develop their own system of promoting the increase of response rates (e.g., adding a percentage point to overall grade, setting aside time in class for students to complete the evaluations; dropping a low homework grade, extra credit points)
9. Consider carefully the use of school-wide incentives. Having lotteries for big-ticket items (e.g., television, iPad, iPhone) may not have the same impact as vouchers for food or bookstore items that everyone can earn.
10. Consider carefully the use of school-wide disincentives such as not permitting students to access grades until they complete the evaluation of that class. Students may simply select numbers on the evaluation form out of anger for being forced to participate, thereby generating evaluation scores that are not valid.

Currently, on-line evaluations appear to be the preferred way to evaluate instruction. While some IHE still use paper and pencil formats, the general trend is to gradually eliminate that format in favor of on-line evaluations. Although there continue to be some disadvantages to the on-line format, most notably the lower response rates, the economic advantage seems to far outweigh the response rate problem in these times of tight budgets and funding shortages. As we have learned in this article, even small attempts to influence response rates cause positive results. Thus, IHE that experience low response rates can try some of the recommendations in this article to bring their on-line response rates closer to those that they enjoyed in the paper and pencil format. Reverting back to the paper and pencil format does not appear to be a viable option.

References

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153-166. doi:10.1023/A:1008168421283

Anderson, J., Brown, G., & Spaeth, S. (2006). Online student evaluations and response rates reconsidered. Innovate, 2(6). Retrieved from http://www.innovateonline.info/index.php?View=article&id=301

Anderson, H.M., Cain, J., & Bird, E. (2005). Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education, 69*(1), 34-43. doi:10.5688/aj690105

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H. & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Electronic Education. 37*(1), 21-38. doi:10.3200/JECE.37.1.21-37

Ballantyne, C. (2003). Online evaluations of teaching: An examination of current practice and considerations for the future. In D.L. Sorenson & T.D. Johnson (Eds.) *Online Student Ratings of Instruction, New Directions for Teaching and Learning,* No.96, Winter 2003 (pp. 103-112). San Francisco: Josey-Bass. doi:10.1002/tl.127

Beran, T. N. & Rokosh, J. L. (2009). Instructors' perspectives on the utility of student ratings of instruction. *Instructional Science, 37,* 171-184. doi: 10.1007/s11251-007-9045-2

Bullock, C. D. (2003). Online collection of midterm student feedback. *New Directions for Teaching and Learning, 96,* 95-103. doi:10.1002/tl.126

Centra, J. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness.* San Francisco, CA: Jossey-Bass Inc Pub.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44,* 495-518. doi:10.1023/A:1025492407752

Cook, C., Heath, F., & Thompson, R. (2000). A meta-analysis of response rates in web- or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821-836. doi:10.1177/00131640021970934

Dommeyer, C. J., Baum, P., Chapman, K. S., & Hanna, R. W. (2002). Attitudes of business faculty towards two methods of collecting teaching evaluations: Paper vs. online. Assessment and Evaluation in Higher Education, 27, 455-462. doi:10.1080/0260293022000009320

Dommeyer, C.J., Baum, P., Hanna, R.W., & Chapman K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment and Evaluation in Higher Education, 29*(5), 611-623. doi:10.1080/02602930410001689171

Fowler, F. J. (2009). *Survey research methods, (4th Ed).* Thousand Oaks: CA: SAGE. doi:10.4135/9781452230184

Gaillard, F. D., Mitchell, S. P., & Kavota, V. (2006). Students, faculty, and administrators' perception of students' evaluations of faculty in higher education business schools. *Journal of College Teaching & Learning, 3,* 77-90.

Gamliel, E. & Davidovitz, L. (2005). Online versus traditional teaching evaluation: Mode can matter. *Assessment and Evaluation in Higher Education, 30,* 581-592. **doi:** 10.1080/02602930500260647

Gigliotti, R. J. & Buchtel, F. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82,* 341-351. doi:10.1037/0022-0663.82.2.341

Greenwald, A. G. & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209-1217. doi:10.1037/0003-066X.52.11.1209

Haefele, D. L. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education, 7*(1), 21-31, doi:10.1007/BF00972346

Heath., N., Lawyer, S., & Rasmussen, E. (2007). Web-based versus paper and pencil course evaluations. *Teaching of Psychology, 34(4)*, 259-261. doi:10.1080/00986280701700433

Hobson, S. M. & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching, 49,* 26-31. doi:10.1080/87567550109595842

Johnson, T. D. (2003). Online student ratings: Will students respond? In D.L. Sorenson & T.D. Johnson (Eds.) *Online Student Ratings of Instruction, New Directions for Teaching and Learning,* No.96, Winter 2003 (pp. 49-59). San Francisco: Josey-Bass. doi: 10.1002/tl.122

Johnson, V. E. (2002). Teacher course evaluation and student grades: An academic tango. *Chance, 15,* 9-16.

Jones, C. (2009). *Nonresponse bias in online course evaluations*, (Doctoral Dissertation, James Madison University). Retrieved from http://search.proquest.com/docview/305166971?accountid=8483

Kulik, J. A. (2001), Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research*, *2001,* 9–25. doi: 10.1002/ir.1

Layne, B. H., DeCristoforo, J., R., & McGinty, D. (1999). Electronic vs traditional student ratings of instruction. *Research in Higher Education, 40*, 221-232.

Liegle, J. O. & McDonald, D. S. (2004). Lessons learned from online vs. paper-based computer information students' evaluation systems. In *The Proceedings of the Information Systems Education Conference.* (Newport): §2214.

Liu, Y. (2006). A comparison of online versus traditional student evaluation of instruction. *International Journal of Instructional Technology and Distance Learning*, *3*(3), 15-30.

Nasser, F. & Fresko, B. (2002) Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education, 27,* 187-198.

Norris, J. & Conn, C. (2005). Investigating strategies for increasing student response rates to online delivered course evaluations. *Quarterly Review of Distance Education, 6*(1), 13-29.

Nulty, D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education*, 33(3), 301-314. doi:10.1080/02602930701293231

Prunty, P. K. (2011, December). Bolstering student response rates for online evaluation of faculty. *E-xcellence in Teaching Essay.*

Ravenscroft, M. & Enyeart, C. (2009). *Online student course evaluations. Strategies for increasing student participation rates*. Washington, DC: The Advisory Board Company http://tcuespot.wikispaces.com/file/view/Online+Student+Course+Evaluations+-+Strategies+for+Increasing+Student+Participation+Rates.pdf

Remedios, R. & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal, 34,* 91-115. doi:10.1080/01411920701492043

Student evaluations of teaching: A comprehensive study of online verses paper modes. (2011, Spring). *San Jose State University.* Retrieved from http://www.oir.sjsu.edu/oirblog

Sax, L., Gilmartin, S., & Bryant, A. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education, 44(4)*, 409-432. doi:10.1023/A:1024232915870

Sheehan, K., & McMillan, S. (1999). Response variation in e-mail surveys: An exploration. *Journal of Advertising Research, 39.*45-54.

Smith, B. (2011). *Improvement of response rates to online measures of instruction: Task Force report.* Manuscript, Office of the Provost, Ball State University, Muncie, IN.

Spencer, K. J. & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment and Evaluation in Higher Education, 27,* 397-409. doi:10.1080/026029302200000928

Sue, V. M., & Ritter, L. A. (2007). *Conducting online surveys.* Thousand Oaks, CA: SAGE.

Thorpe, S. W. (2002). *Online student evaluation of instruction: An investigation of non-response bias.* Paper presented at the 42nd Annual Forum for the Association for Institutional Research. Toronto, Ontario, Canada.

Acknowledgement

I appreciate the assistance of the Task Force on Online Teaching Evaluations at Ball State University who compiled much of the literature used in this chapter. Task Force members were Brien Smith, Chair, Cheryll Adams, Carol Friesen, James Jones, Rai Peterson, and Carolyn Walker.

Contact Information

Contact Cheryll Adams at cadams@bsu.edu

What's the Story on Evaluations of Online Teaching?

Michelle Drouin
Indiana University–Purdue University Fort Wayne

The Back Story

During the past decade, there has been a great shift in higher education towards online learning. According to recent statistics from the Pew Research Center, more than three-quarters of the colleges and universities in the United States now offer online courses, and 23% of all college graduates and 46% of recent graduates (from last decade) reported taking at least one online class (Parker, Lenhart, & Moore, 2011). Moreover, in line with the trend of enrollments in online courses increasing at a greater rate than enrollment in higher education overall, 50% of college presidents predict that most students will be taking classes online 10 years from now (Parker et al. 2011). This drastic increase in online course offerings has spurred many questions among educators about the equivalency of online and face-to-face (f2f) courses.

From a functional perspective online and f2f courses may not be so different, at least in how they are perceived. A recent survey from the Pew Foundation showed that approximately half (51%) of college presidents and 29% of the general public believe that online courses offer the same education value as f2f courses (Parker et al., 2011). However, online courses often have very different structures than f2f courses and have their own models for effectiveness (e.g., Peltier, Schibrowsky, & Drago, 2007). These structural differences are often motivated by the lack of physical presence in online courses, which obliges online instructors to use other means to foster students' feelings of connectedness to the course. For example, online courses might include more social networking tools (e.g., wikis, blogs, instant messaging, social networking sites) to enhance students' social presence (e.g., Joyce & Brown, 2009). Additionally, online instructors might use more prompt and effective communication methods, as timely instructor feedback appears to be a key feature for fostering student satisfaction in online courses (Moore, 2005). Each of these communication strategies (social networking and instructor feedback) might be used to enhance students' connectedness to their courses, which has been shown to bolster online students' achievement and satisfaction (e.g., Richardson & Swan, 2003). Although social interactions are also a key feature of effective f2f courses (Chickering & Gamson, 1987), instructors would necessarily employ different strategies to foster social interactions in f2f vs. online courses. This single example highlights an important point—one which inspired the writing of this chapter: instructors often employ different pedagogical techniques in their online and f2f courses. Because of these differences, the evaluation strategies used to assess f2f courses are not necessarily relevant to the online setting.

Unfortunately, the shift in higher education towards online courses happened so quickly that little early attention was devoted to the evaluation of online courses (Loveland, 2007). Instead, instructors were likely focused on learning the online course technologies and adapting their pedagogical strategies to the online environment; they therefore either skipped student and peer evaluations in online courses or used the same methods that were used in f2f courses (Compora, 2003). However, instructors soon learned that traditional methods of course evaluation did not adequately measure the instructional design and delivery methods of online courses (e.g., Harrington & Reasons, 2005; Loveland, 2007). Therefore, evaluations of online teaching began to surface, aimed specifically at evaluating online courses.

The Present Story

Today, evaluations of online teaching (EOTs) often take the form of general course rubrics, which can be used for peer evaluations or by instructors as checklists in their own course design. These rubrics originated from a variety of sources, including non-profit organizations aimed at enhancing online education, such as Maryland Online (Quality Matters), the Illinois Online Network (Quality Online Course Initiative; QOCI), and the Monterey Institute (Online Course Evaluation Project; OCEP). Universities have also developed general rubrics for widespread use, such as the Online Course Assessment Tools (OCATs) from Texas A&M and Western Carolina University and the self-assessment Rubric for Online Instruction (ROI) developed by California State University-Chico. General EOTs have usually been created by a team of contributors, who have developed the rubrics based on reviews of the online education literature and their own experiences in designing and reviewing online courses. All of these general EOT rubrics are available online, mostly in .pdf format, with the exception of the Texas A&M rubric, for which those outside of the university system must register for free online rubric access (more detailed descriptions below).

Unfortunately, much less attention has been devoted to student evaluations of online teaching (SEOTs), and only a few studies have produced SEOT rubrics. Some early versions of SEOTs were simply standard f2f rubrics appended with questions about the online environment. For example, Tallent-Runnels et al. (2005) added a "technology evaluation" portion to an existing f2f survey so that they could examine the technological features of their online course. However, a few student evaluations have been created specifically for online courses, including the Student Evaluation of Web-Based Instruction (SEWBI; Stewart, Hong, & Strudler, 2004), the Student Evaluation of Online Teaching Effectiveness (SEOTE; Bangert, 2008), and a more recent online course assessment tool created by Rothman, Romeo, Brennan, and Mitchell (2010). None of these SEOTs are available as online rubrics; instead, they can be found within the articles cited.

Before discussing the specific rubrics that might be most useful to online instructors, it is first important to discuss their commonalities. This discussion naturally begins with a framework, and the framework that is most relevant to the current topic is one related to the effectiveness of online courses. A number of researchers have studied the structural components related to effective online course delivery and have, in turn, developed various models of effective online teaching. In one of the more comprehensive studies, Peltier et al. (2007) examined several of these models and distilled them down to six factors that contributed to students' perceptions of online teaching effectiveness: "(a) student-student interactions, (b) student-instructor interactions, (c) instructor support and mentoring, (d) lecture delivery quality, (e) course content, and (f) course structure" (p. 141). They then used structural equation modeling to determine whether each of these factors contributed significantly to students' perceived quality of their online courses; each did so either directly or indirectly through one of the other factors.

Using Peltier et al.'s (2007) model as a framework, I have summarized the key features of the most widely-cited general (peer and self) and student EOTs in Table 1. As can be seen, the rubrics vary somewhat in scope and length; however, all address most or all of the essential factors in Peltier et al.'s model. Therefore, each is relevant to the current online pedagogical research and therefore potentially useful for the formative and summative feedback it can provide.

Table 1

Summary of Key Features of Popular Self, Peer, and Student Evaluations of Online Teaching (EOTs)

	Type of Evaluation	Total Categories /Items	Student-Student Interactions	Student-Instructor Interactions	Instructor Support and Mentoring	Lecture Delivery Quality	Course Content	Course Structure
Quality Matters; Maryland Online	Peer & Self	8/41	Learner Interaction and Engagement		Assessment and Measurement; Learner Support	Course Technology	Learning Objectives (Competencies); Instructional Materials	Course Overview and Introduction; Accessibility
QOCI; Illinois Online Network	Peer & Self	6/82	Communication, Interaction, & Collaboration		Learner Support & Resources	Instructional Design	Student Evaluation and Assessment	Web Design
OCEP; Monterey Institute	Peer & Self	8/52	Communication Tools and Interaction		Assessments and Support Materials	Course Features and Media Values	Scope and Scholarship; Developer Comments	Course Developer and Distribution Models; User Interface; Technology Requirements and Interoperability

OCAT; Texas A&M	Peer & Self	8/89	Communication & Collaboration	Syllabus & Other Documents	Learning Outcomes & Activities	Content; Assessment & Evaluation	Navigation & Organization; Consistency; Accessibility
OCAT; Western Carolina	Peer & Self	5/67	Learner Interaction	Learner Objectives & Competencies		Resources & Materials; Learner Assessment	Course Overview & Organization
ROI; CSU Chico	Self & Awards	6/25	Instructional Design & Delivery; Faculty Use of Student Feedback	Learner Support & Resources	Innovative Teaching with Technology	Assessment & Evaluation of Student Learning	Online Organization & Design
Rothman et al., 2010	Student	6/25	Instructor Feedback and Communication	Clarity of Outcomes and Requirements	Technological Tools; Content Format	Appropriateness of Readings and Assignments	Course Organization
SEOTE (Bangert, 2008)	Student	7/26	Cooperation Among Students; Student-Faculty Contact	Prompt Feedback		Active Learning; High Expectations	Time on Task; Diverse Talents and Ways of Learning
SEWBI (Stewart, Hong, & Strudler, 2004)	Student	8/59	Instructor and Peer Interaction	Class Procedures & Expectations	Online Applications	Content Delivery	Appearance of Web Pages; Hyperlinks & Navigation; Technical Issues

How Do I Choose the Right Story?

Clearly, a variety of choices exists for evaluating online courses. With regard to self- and peer-EOTs, the rubric you choose may depend on your institutional policies, as some institutions, like mine, have preferred peer online evaluation rubrics. However, SEOT choices are left typically to departments or individual instructors; thus, there is likely to be more flexibility in choosing or developing an SEOT that meets your course evaluation goals. In either case, the EOT that is most appropriate for your course may depend on the teaching style you have adopted within the course.

Anderson and Dron (2011) summarized the different teaching approaches that have been used in distance education during its brief three-generational history. According to these authors, technological and theoretical developments in the past decades have influenced instructors' designs of their distance education courses. Early courses were based on co*gnitive-behaviorism*, where learners interacted individually with materials. Later, there was a shift towards *social constructivism*, which emphasized learners' interactions with the instructor and other learners to create knowledge. More recently, distance education instructors have begun to adopt *connectivist* approaches, where learners interact with people and content through Web-based networks with a goal of finding and applying knowledge. Some of the unique design components of these pedagogical approaches to distance education are summarized in Table 2.

Table 2

Summary of Three Generations of Instructional Approaches in Online Education as Detailed by Anderson and Dron (2011)

Pedagogical Approach	**Content delivery method**	**Student learning method**	**Teaching presence**	**Interactions within course**
Gen. I: *Cognitive-Behaviorism*	Print (including correspondence), video, audio files	Student interacts with material	Possibly communicated through correspondence, and indirectly through audio, and video	Student has freedom to interact with content; no or few S-S or S-I interactions are required or encouraged
Gen. II: *Social-Constructivism*	Interactions via audio, video, and print (group and one-to-one interactions)	Student constructs own knowledge through interactions	Teacher actively guides students through tasks; provides direct instruction and initiates discussions	Student interacts actively with content, instructor, and other students

Gen. III: *Connectivism*	Networks of resources (people, digital objects, and content on the Web)	Student finds and applies knowledge through networks; little emphasis on memorization	Teacher creates and sustains networks (e.g., wikis, threads, Twitter)	Student interacts through networks of current and archived resources

EOT recommendations for different pedagogical approaches

The *cognitive-behavioral* approach was popularized at a time when interactivity with students at a distance was difficult because of the limited interactive technologies available (Anderson & Dron, 2011). However, this approach is still utilized today, and it has been enriched by the capability of presenting learning materials via multimedia (using voice and image of instructor) and personalized correspondence from instructors. That said, this approach is based upon individualized learning, and the student should be able to learn and demonstrate learning simply through interaction with course materials (Anderson & Dron, 2011). Therefore, the evaluations that would be most relevant to this pedagogical approach would place more emphasis on the efficiency of delivering the content (e.g., organization of the course, navigation, etc.) and less emphasis on student-student interactions and collaborative learning (see Table 3). With regard to peer and self EOTs, the Quality Matters, Western Carolina, and Texas A&M rubrics contain many criteria related to course structure and navigability and fewer related to student collaboration. Meanwhile, the SEWBI and Rothman et al. (2010) student EOTs are also geared more towards structure and navigation. However, it must be noted that with some rubrics (e.g., Quality Matters), the instructor would "lose points" when features (e.g., collaboration tools) were not used within the course. Consequently, instructors using the cognitive-behavioral approach (or any approach) might also want to use customizable peer review form like the one created by Wood and Friedel (2009). Their Peer Review of Online Learning and Teaching system allows peer reviewers to choose the questions that are most relevant to the course they are reviewing (from a large bank) or even create their own questions.

With the *social-constructivis*t approach, students create their own knowledge through active learning experiences and collaborative learning. This is one of the more popular styles of online teaching today and has long been endorsed in f2f education, as evidenced by the much-cited "Seven Principles for Good Practice in Undergraduate Education" (Chickering & Gamson, 1987). In fact, a number of researchers (e.g., Bangert, 2008; Gorsky & Blau, 2009; Weiss, 2010) have suggested that Chickering and Gamson's principles be used as a baseline for the development of good online courses. Accordingly, these principles have been used as a framework to examine the different evaluation categories of some of the more popular, general EOTs (e.g., Weiss, 2010). Instructors employing a constructivist strategy might use the class and group discussion tools (e.g., discussion boards, wikis, and blogs) to help students construct knowledge in the course. They might also use interactive exercises (e.g., quizzes, flashcards, and crossword puzzles) so that students have independent active learning opportunities. Consequently, evaluations of the collaborative, communicative, and active learning tools might provide the most useful feedback for instructors using a social-constructivist approach (see Table 3). In terms of peer and self EOTs, the Quality Matters (QM) rubric or the QOCI might be most useful for those employing a constructivist design. The QM rubric focuses specifically on active learning and has an entire category devoted to learner interaction and engagement. Meanwhile, the QOCI has a category devoted to communication, interaction, and collaboration and also has specific criteria on student-student interaction, creating forums for community, and group work. With regard to SEOTs, Bangert's (2008)

rubric (SEOTE) was designed specifically to address Chickering and Gamson's seven principles; thus, it would be the most relevant SEOT for those using a constructivist approach in their online courses.

The newer *connectivist* approach also relies heavily upon communication and interactions, but in this case, the instructor's responsibility is to create networks in which for these interactions to occur. The learner then creates his or her own knowledge through connecting with resources, and communicates any knowledge gained through other learning networks (e.g., blogs, wikis, or social networking posts). This creates an extended, collaborative community of learners where the instructor acts as a guide or role model in the connectivist process (Anderson & Dron, 2011). For instructors using the connectivist approach, rubrics that evaluate the technological aspects of the course, such as its navigability and the technological tools employed, might provide the most useful feedback (see Table 3). With regard to peer and self EOTs, Quality Matters is a well-rounded rubric with specific criteria related to the accessibility of course tools, and the OCEP (Monterey Institute) and the ROI (CSU Chico) both provide a good set of criteria that evaluate the more technological aspects of the course. The SEOTs that might be most useful for instructors using the connectivist approach are the SEWBI (Stewart, Hong, & Strudler, 2004), which focuses specifically on the navigability of the site and online applications, and the rubric developed by Rothman et al. (2010), which includes criteria on organization, formatting, and technological tools. However, none of these is geared specifically towards the connectivist approach; thus, there is a need for more research in this area.

Table 3

Summary of Recommendations for EOTs Based on Pedagogical Approach

Approach	**Peer and Self EOT**	**Student EOT**
Cognitive Behavioral	Quality Matters (peer & self); Western Carolina OCAT (peer & self); Texas A&M OCAT (peer & self); ROI (self);	SEWBI; Rothman et. al evaluation
Social-Constructivism	Quality Matters (peer & self); QOCI (peer & self)	SEOTE
Connectivism	Quality Matters (peer & self); OCEP (peer & self); ROI (self)	SEWBI; Rothman et al. evaluation

In the End, Isn't this All Just the Same Story?

Although the rubrics do vary slightly in the instructional features they assess, in the end, as Table 1 demonstrates, they do have a lot of commonalities. Therefore, you might be wondering how to choose one rubric over another, especially when they seem to be assessing similar qualities of online teaching. To this, I would answer that the review processes, specifically the processes related to general EOTs, vary greatly. In the list below, I summarize the defining features of the review process for each of peer/self rubrics presented in Table 1.

Quality Matters. Quality Matters (QM) is probably the most widely-known online course review program. The QM rubric is available free online http://www.qmprogram.org/rubric and provides

detailed examples (annotations) of the types of materials and features that could be used to meet criteria. For those wishing to have their courses officially QM certified, they must undergo a rigorous peer review process completed by team of three trained reviewers, one of which is external to the institution ($1000 fee through QM; institutions on full subscriptions manage their own reviews). QM also trains peer reviewers through an online course, and these trainings are usually arranged by academic institutions, which must subscribe (there is a fee involved) to the QM higher education program (the 500+ subscribers are listed here: http://www.qmprogram.org/qmresources/subscriptions/state.cfm). Only peer reviewers who complete the training are certified to conduct official peer reviews. The reviewer completes the rubric by awarding points (0-3 depending on the criterion) for standards being assessed. In order for a course to be approved, it must meet all essential (3-point) standards as well as attaining a total score of 85%.

Illinois Online Network: QOCI. The Illinois Online Network (ION) provides online learning resources to more than 30 academic institutions nationally (http://www.ion.illinois.edu/partners/nationalpartners/index.asp). Again, the rubric (brief and annotated versions) is available free online http://www.ion.uillinois.edu/initiatives/qoci/rubric.asp. This rubric is free to be copied, distributed, and adapted through the Creative Commons License for those wanting to conduct peer reviews or self-assessments. An official QOCI review can also be arranged (for approximately $200-$250 per course), which entails a one-person review of the design and construction of the course. The rubric criteria are straightforward and reviewers provide feedback on each criterion on a 4-point scale (*non-existent*, *developing*, *meets*, or *exceeds standards*; there is also an n/a option).

Monterey Institute: OCEP. Monterey Institute provides free online access to its evaluation rubric http://www.montereyinstitute.org/pdf/OCEP%20Evaluation%20Categories.pdf, which has straightforward criteria, definitions, and goals stated. However, the evaluative review process, which consisted of a team of reviewers: a subject matter expert, an online multimedia professional, and a technical consultant, is no longer active.

Texas A&M: OCAT. The Texas A&M rubric is the only rubric in this list where the evaluation is completed online https://elearningtools.tamu.edu/checklist/login.do, within a system that requires registration (free). The rubric consists of a checklist, and if the course being evaluated has the criterion listed (details are available by hovering mouse over the criteria), then you check the corresponding box. After the evaluation is submitted, the system generates an immediate, online assessment report, which gives a point value and percentage for each of the criteria and an overall total. This evaluation can be printed.

Western Carolina: OCAT. Western Carolina's rubric has clear, straightforward criteria and is also available free online http://www.wcu.edu/WebFiles/PDFs/facultycenter_OCAT_v2.0_25apr07.pdf. Reviewers provide feedback on each criterion by indicating whether the feature is evident or not evident (an n/a option is also available). The rubric also includes a section where the instructor can respond to the assessment. This might be especially useful for those intending to use the rubric for formative assessment.

California State University (CSU)-Chico: ROI. The CSU-Chico rubric http://www.csuchico.edu/tlp/resources/rubric/rubric.pdf is unique in this group because it is not intended for peer review. Instead, it was developed for instructors to use during course design and for self-assessments. More than 100 campuses have inquired about using this rubric, and the institution has given permission for anyone to adapt it to suit their needs (through the Creative Commons License). In comparison to other rubrics in this category, this rubric is incredibly straightforward because it

provides specific explanations for what would be considered *baseline*, *effective*, and *exemplary* for each criterion. It is also unique because can also be used by instructors within the CSU-Chico system to gain recognition for exemplary online courses. This model (using a rubric to recognition of exemplary online courses) is a promising one for institutions striving to maintain quality in their online education courses.

Although the SEOTs (student evaluation rubrics) are not commercialized and therefore do not have a formal review process associated with them, a short description of their defining features might also be useful to those wishing to adopt an SEOT.

Rothman et al. (2010) Rubric. The Rothman et al. rubric provides a fairly broad overview of course components and includes both general criteria (e.g., "Assignments were appropriate and effective for learning course content.") and specific criteria (e.g., "Dates on the course schedule corresponded to drop box and discussion board submissions," and "Online course materials were free of spelling errors and grammatical errors."). For all questions, students indicate their agreement with statements on a Likert scale from 1 = *strongly disagree* to 5 = *strongly agree.* One distinct advantage of the Rothman et al. rubric is that this SEOT has high reliability (as noted by the authors), and the authors also indicate specific areas where student ratings tend to be lower (e.g., technology and instructor feedback).

Student Evaluation of Online Teaching Effectiveness (SEOTE). This rubric, developed by Bangert (2008), is unique because it was designed specifically to align with Chickering and Gamson's (1987) seven principles. Therefore, many of the items are applicable in both f2f and online courses. For example, under the category of "Prompt Feedback," the rubric has a criterion that states "My questions about course assignments were responded to quickly" (p. 37). For all questions, students indicate their agreement with statements related to these seven principles on a Likert scale from 1 = *strongly disagree* to 6 = *strongly agree.* This rubric has two distinct advantages: (1) it provides an instant theoretical framework for your course evaluations, and (2) it could be tailored for use in both f2f and online courses so that f2f and online evaluations were comparable.

Student Evaluation of Web-Based Instruction (SEWBI). The SEWBI (Stewart, Hong, & Strudler, 2004) is more than twice as long as either of the other SEOTs, but it has only two more evaluation categories than the Rothman et al. (2010) rubric and only one more evaluation category than the SEOTE. This rubric places a strong emphasis on the technological aspects of the course, like the course tools and web pages. For example, the rubric asks for students to evaluate the ease of use of ten different course tools (e.g., video player, chat rooms, simulations) and also has eight questions about the appearance and ease of use of web pages. For most questions, students indicate their agreement with statements on a Likert scale from 1 = *strongly disagree* to 5 = *strongly agree.* Overall, this rubric would be most useful for instructors who use a lot of multimedia and want to focus on evaluating the technological aspects of their course.

Yes, It Might Just Be the Same Old Story

Although I have just presented rather compelling evidence that these EOT rubrics are different, their differences relate mostly to their processes. In fact, their components are quite similar, just as the title of this section suggests. For those hoping to incorporate the different features of several rubrics to design a high quality online course, there are a number of specific criteria that appear on all (or nearly all) of the rubrics, both general (peer and self) and student.

1. *Student-Student and Student-Instructor Interactions*—Does the course allow ample opportunity for students to interact with the instructor and other students? Does the instructor communicate clearly and regularly?
2. *Instructor Support & Mentoring*—Are the learning objectives clear and aligned with course description? Does the instructor provide links for relevant software, media, and external resources?
3. *Lecture/Content Delivery Quality*—Does the course use a variety of technological tools and multimedia elements that address diverse ways of learning? Are these tools readily available?
4. *Course Content*— Is the course material relevant to the course level and specific learning objectives? Are the assessments and assignments aligned with the learning objectives? Do students have an opportunity to get feedback on their work? Are multiple types of assessment of student learning used?
5. *Course Structure*—Do students know where to start? Whether topical or chronological, is the course organized logically? Is navigation clear? Is it user-friendly?

These criteria convey some best practices for online education, but they could also serve as recommendations for best practices in the f2f classroom. In fact, Wang, Dziuban, Cook, and Moskal (2009) showed that whether classes were in the f2f or online format, an instructors' overall rating was dependent on four of these five factors: their communication of ideas and information, their facilitation of learning, assessment of student progress, and organization of the course. This suggests that the online and f2f teaching environments are not actually as different as they may seem. Although they may have different technological components, and the effectiveness of these technological components should be measured, similar factors contribute to positive student ratings in both environments.

References

Anderson, T., & Dron, J. (2011). Three generations of distance education pedagogy. *International Review of Research in Open and Distance Learning*, *12*, 80–97.

Bangert, A. W. (2008). The development and validation of the Student Evaluation of Online Teaching Effectiveness. *Computers in the Schools, 25*, 25–47. doi: 10.1080/07380560802157717

Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *American Association of Higher Education Bulletin*, *39*, 3–7.

Compora, D. (2003). Current trends in distance education: An administrative model. *Online Journal of Distance Learning Administration*, *6.* Retrieved from http://www.westga.edu/~distance/ojdla/summer62/compora62.html

Harrington, C. F., & Reasons, S. G. (2005). Online student evaluation of teaching for distance education: A perfect match? *The Journal of Educators Online*, *2*(1), 1–12.

Joyce, K. M., & Brown, A. (2009). Enhancing social presence in online learning: Mediation strategies applied to social networking tools. *Online Journal of Distance Learning Administration*, *12*(4). Retrieved from http://www.westga.edu.ezproxy.lib.ipfw.edu/~distance/ojdla/winter124/joyce124.html

Loveland, K. A. (2007). Student evaluation of teaching (SET) in web-based classes: Preliminary findings and a call for further research. *The Journal of Educators Online*, *4*(2), 1–18.

Moore, J. C. (2005). The Sloan Consortium quality framework and the five pillars. *The Sloan Consortium*. Retrieved from www.sloanconsortium.org/publications/books/qualityframework.pdf

Parker, K., Lenhart, A., & Moore, K. (2011, August 28). The digital revolution and higher education. *Pew Internet and American Life Project*. Retrieved from http://www.pewinternet.org/Reports/2011/College-presidents/Summary.aspx?view=all

Peltier, J. W., Schibrowsky, J. A., & Drago, W. (2007). The interdependence of the factors influencing the perceived quality of the online learning experience: A causal model. *Journal of Marketing Education*, *29*, 140–153. doi: 10.1177/0273475307302016

Richardson, J. C., & Swan, K. (2003). Examining social presence in online courses in relation to students' perceived learning and satisfaction. *Journal of Asynchronous Learning Networks, 7,* 68–88.

Rothman, T., Romeo, L., Brennan, M., & Mitchell, D. (2010). 21st century best practice and evaluation for online courses. *International Conference: The Future of Online Education*. Retrieved from: http://www.pixel-online.net/edu_future/common/download/Paper_pdf/ELE02-Rothman,Romeo,Brennan,Mitchell.pdf

Stewart, I., Hong, E., & Strudler, N. (2004). Development and validation of an instrument for student evaluation of the quality of Web-based instruction. *The American Journal of Distance Education, 18,* 131–150.

Wang, M. C., Dziuban, C. D., Cook, I. J., Moskal, P. D. (2009). Dr. Fox rocks: Using datamining techniques to examine student ratings of instruction. In M. C. Shelley II, L. D. Yore, & B. Hand (Eds.), *Quality research in literacy and science education: International perspectives and gold standards*, 383–398. Dordrecht, Netherlands: Springer.

Weiss, A. (2010). Comparison of rubric categories and seven principles. Retrieved from https://apps.lis.illinois.edu/wiki/display/elearning/Online+Course+Evaluation+Rubrics

Wood, D., & Friedel, M. (2009). Peer review of online learning and teaching: Harnessing collective intelligence to address emerging challenges. *Australasian Journal of Educational Technology*, *25*, 60–79.

Contact Information

Michelle Drouin
Department of Psychology
2101 E. Coliseum Blvd.
Fort Wayne, IN 46835
drouinm@ipfw.edu

Using Course Portfolios to Assess and Improve Teaching
Paul Schafer, Elizabeth Yost Hammer, Jason Berntsen
Xavier University of Louisiana

Using Course Portfolios to Assess and Improve Teaching

For many college teachers, the word assessment is met with general skepticism. We often associate it with tedious accreditation reports and time-consuming busy work. However, when we take a step back and examine the bigger picture, we should be able to embrace assessment (especially assessment of teaching) for the valuable data it can provide. After all, as faculty, we demand that students "show us the data" and provide evidence to support their claims. How can we then feel comfortable with claiming "I am a good teacher" with no evidence to support this assertion aside from a gut feeling? Further, how can we improve on any weaknesses if these areas are left unidentified? Those of us who are truly interested in improving our teaching practice and enhancing student learning should be holding ourselves to a high standard and encouraging, not resisting, meaningful assessment. In fact, Brookfield (2006) writes, "the essence of skillful teaching lies in the teacher constantly researching how her students are experiencing learning and then making pedagogical decisions informed by the insights she gains from student responses" (p. xi).

Luckily, as this book illustrates, there are many evaluation techniques at our disposal, and we can choose techniques that best match our courses, disciplines, institutional cultures, and individual teaching styles. In this chapter we will describe one such technique: the course portfolio. In 1997, Xavier University of Louisiana developed systematic, campus-wide course portfolio working groups (CPWGs). This initiative offers an effective faculty-driven means of assessing and improving teaching. It has helped scores of individual faculty participants improve their courses, and has contributed to the cultivation of a culture of teaching excellence on our campus. In what follows, we will offer a few background words about assessment and the different types of portfolios, before turning to the details of Xavier's program. We conclude with some practical advice for individuals and institutions who are interested in utilizing course portfolios.

Assessment and Portfolios

Experts talk about the distinction between formative and summative assessments (e.g., Cerbin, 1994; Taras, 2005). In a faculty context, summative assessment includes techniques whose sole purpose is to *evaluate* one's teaching (typically, end-of-term student evaluations). Summative assessment is often used for personnel decisions such as tenure and promotion. In contrast, the purpose of formative assessment is to solicit feedback in order to *improve* one's teaching. Gathering student feedback on an on-going basis allows teachers to adjust their methods to address the needs of the class, ultimately enhancing student learning.

One benefit of a portfolio approach to assessment is that it can provide both summative and formative data. With regard to summative assessment, portfolios allow for a more comprehensive and complex picture of teaching than a single end-of-term survey; with regard to formative assessment, the portfolio allows for continuous, systematic feedback to which the teacher can respond. Such portfolios are works in progress that allow the instructor to adjust teaching methods and redesign aspects of a course after examining results of both student and self-learning.

As this analysis suggests, the portfolio can take multiple forms. It can be an entirely personal document, an instrument of summative assessment, or it can be utilized as a contribution to the growing scholarly literature on teaching and learning. A portfolio can cover the full range of one's teaching experiences or focus on a single course. The portfolio can also be used to trace the development of one's teaching over time. In this way, the concrete instances of teaching are not only placed in the context of a general vision of teaching but also in a historical context, thereby forming a narrative. As a finished product, the portfolio can take the traditional form of hard copies collected in a binder, or it can be an electronic portfolio or e-portfolio (Gross, 2009).

Portfolio advocates suggest that their greatest value lies in the reflection on teaching that is intrinsic to the portfolio experience, and suggest that this benefit is maximized when one creates a portfolio with others, as in the context of a portfolio working group (Hutchings, 1996). Of course, merely collecting teaching materials and other teaching-related documents in a single place will not bring about this reflection. Thus, whatever form the portfolio takes, it should be structured and selective (Edgerton et al., 1991). Besides facilitating reflection, a structured and selective portfolio is more likely to be effective for communicating one's teaching practices to others.

A related benefit of portfolios lies in their ability to provide a richer portrayal of teaching than other means of assessment. This is particularly true of portfolios that combine primary documents with personal reflections. Faculty members do not work in a vacuum; they teach particular things to particular students at particular times and particular places. Portfolios that contain only personal reflections can fail to effectively capture these concrete particulars. Conversely, a collection of primary documents without annotation or accompanying reflective essays will not reveal how the concrete particulars relate to each other or put them in the context of a general vision. However, when a portfolio combines artifacts of teaching with reflective commentary, it captures particular instances of teaching in a way that reveals the general approaches to pedagogy that help give them meaning (Edgerton et al., 1991).

Another benefit of portfolios, one emphasized by the pioneers of their use, is that they can make acts of teaching community property (Hutchings, 1996). Many have pointed out that designing and implementing a course can be approached as a scholarly endeavor, as opposed to something that teachers do *in addition* to scholarship (Boyer, 1990; Hutchings and Shulman, 1999). However, as Lee Shulman (1998) notes, scholarship is by definition public, open to review, and available for one's professional peers to use and build upon. Portfolios thus have the potential to transform teaching from a scholarly activity into a genuine form of scholarship, and to encourage teachers to view themselves as a forming a community of scholars. This benefit of portfolios fits most squarely, of course, with portfolios created as potential contributions to the literature. However, even a portfolio intended as a private document or instrument of summative assessment can promote an environment of teaching scholars if it is created with others in a working group.

Types of Portfolios

As mentioned above, a portfolio can cover the full range of one's teaching experiences or address a single course. That is, a portfolio can be either a teaching portfolio or a course portfolio. There is considerable overlap between these two types of portfolio. Indeed, a teaching portfolio can be approached as essentially a collection of course portfolios (Gross, 2009), and the course portfolio can be viewed as a teaching portfolio applied to a particular course (Cerbin, 1994).

The obvious comparative advantage of the teaching portfolio is its breadth. It addresses one's teaching over a wider range of time and educational contexts (Hutchings, 1998). With this breadth comes the promise of integration, of assimilating what may sometimes seem like very disparate experiences into a single narrative. There are also, however, important advantages to doing a course portfolio. First, it is at the level of the individual course where most teaching and learning occurs. If, as William Cerbin (1995) argues, the goal of assessment is to investigate the relationship between learning and acts of teaching with the ultimate aim of discovering more effective ways of promoting learning, then the course is the natural unit of analysis for assessment. Where teaching portfolios are weighted toward examining teaching and student portfolios are organized around learning, the course portfolio keeps the focus on the relationship between the two (Cerbin, 1995). A related practical advantage is that creating a course portfolio provides an occasion to reflect on how well the individual elements of a course hang together to form a coherent learning experience (Hutchings, 1998).

Second, the course portfolio is more "portable" than the teaching portfolio. Pat Hutchings (1998) puts the point nicely: "I may or may not be interested in knowing about a colleague's teaching practices in general (which is what I am likely to find in a teaching portfolio), but I might very well be interested in her experience with a course that I myself sometimes teach, or that I rely on as a foundation for one of my own or attempt to build on" (p. 14).

A third advantage, which like the last is related to the power portfolios have to make teaching community property, as well as to the connected point that portfolios can be scholarly documents, is that courses in their specificity are more felicitous objects of scholarly investigation than a faculty member's entire teaching career. As Cerbin (1995) puts it, "[a] course is a coherent unity with specific goals, content, methods, results, and outcomes. In this sense, a course is the pedagogical equivalent of a significant piece of research or scholarship" (p. 4).

A Model Program

Xavier University's Center for the Advancement of Teaching implemented its Course Portfolio Working Group (CPWG) initiative to support faculty in a year-long project that would result in the creation of a course portfolio. The portfolio would then serve as a foundational document upon which the faculty member could build as he or she developed, thus encouraging continued innovation and experimentation with teaching. The program was immediately successful, and we now approach fifteen years of course portfolio groups at Xavier.

The Center's work in this area was initially funded by a generous grant from the Bush-Hewlett Foundation. However, Xavier views the CPWG as a vital part of faculty development, so much so that the university now fully sustains the program. In fact, new faculty members with no prior teaching experience are given a course reduction in their first year in exchange for participation. As a result, the Center works closely with the administration to make sure this initiative is responding to the needs of the university. Recently this has led to the development of thematic portfolio groups, a topic we will discuss below. At this point it is sufficient to note that by addressing the needs of our institution, we have managed to keep this initiative fresh and vibrant over the years.

In its current form, the initiative works as follows. At the beginning of each new academic year the Center sends out a call for participation to the entire Xavier faculty, and the CPWG is called to life once again. What is promised is a year-long exploration of teaching that strives to blend the scholarly and the practical. Participants are offered a small compensation for the time they will be required to commit to the group (currently a stipend for course-related expenses). The aim of the group, on the one hand, is to

encourage faculty participants to make their courses into an object of scholarly inquiry and, on the other hand, to help them make course improvements and become better teachers. The effort of everyone involved is directed towards the elusive goal of making learning visible – both so that we can understand how and why learning happens, but also so that we can maximize the learning that occurs in our classes. At the end of the year, each faculty participant will have a course portfolio that in some way captures and documents what she has done.

Faculty members from across the university answer the call. We encourage participation from all disciplines, not only for the sake of interdisciplinary diversity, but so that each faculty participant may gain a broader perspective on the teaching and learning that happen across the campus. Unlike course portfolio programs at some other institutions, we encourage junior faculty, including first-time teachers, to participate in the group. In the same way that instructors in the Humanities may gain insights from learning first-hand about the challenges faced by Pharmacy instructors, so too may junior faculty learn from their senior colleagues, and vice versa. The resulting working group is a genuinely peer-review-of-teaching instrument: we learn from, assist, and evaluate each other. It is important to emphasize that such peer collaborations must take place in an atmosphere of mutual trust and respect. It is our policy to keep the door closed and to enforce a safe and confidential environment for all meetings.

The CPWG meets a couple of times a month throughout the year. In the fall semester we focus primarily on course reflection and planning, while the spring semester is devoted to putting course changes into practice and assessment. Typically the CPWG opens with a session or two on the scholarship of teaching movement, where we discuss how teaching can be approached in a scholarly way. Ernest Boyer's (1990) influential *Scholarship Reconsidered* is a useful guide for this purpose. We then turn to our own courses and pursue a series of discussions and activities that revolve around three central elements: learning outcomes, teaching practices, and assessment.

The first major course-related exercise taken up in the CPWG focuses on student learning outcomes. Faculty participants write a short memo in which they identify and reflect on the primary learning outcomes for their chosen course. The idea here is to take a closer and more serious look at the course goals, to move past the sort of platitudes that sometimes appear on a course syllabus and to identify the underlying outcomes that we really want our students to achieve. This exercise generates plenty of fruitful discussion during meetings, as faculty initiate the process of scholarly reflection and begin to identify areas of their courses that might be improved.

The second major exercise of the CPWG focuses on teaching methods and practices. Participants again write a short memo, the aim of which is to identify and reflect on the primary instructional elements of the course. Typically, this will include the various assignments, materials, and activities that comprise the course. The peer review aspect shines brightly during this phase of the CPWG as we share and evaluate the everyday teaching practices in our courses. At this point, faculty are encouraged to reflect on the extent to which their current instructional practices really assist students in achieving the course learning outcomes identified in the earlier exercise. As we approach the end of this phase of the project, participants draw up a modified plan for their courses, one that reflects the learning outcomes and instructional practices identified during the peer review process of the working group.

The final major component of the CPWG involves assessment, which is taken up during the last meeting of the fall semester, and continues to be a focal point during the spring. Here again faculty participants write a short memo, only this time they are asked to draft an assessment plan that will allow them to measure the extent to which the soon-to-be introduced course modifications succeed. Not surprisingly,

this tends to be the most difficult and challenging aspect of the CPWG - it is much easier to assign students a grade than to assess the effectiveness of our teaching interventions over the course of a semester! This makes the peer review process all the more important and we devote the bulk of the spring semester to helping each other sort through what is and what is not working, providing real time feedback, encouragement, and support. More details on assessment will be provided in the following section.

At the end of the academic year, each CPWG participant submits a finished course portfolio, which constitutes their successful completion of the program. The completed portfolio is both a warehouse for course and teaching artifacts, as well as a documentation of the scholarly investigation of the course. Portfolio contents vary depending on the nature of the project taken up, but will typically include the following items: table of contents, purpose statement, teaching philosophy, syllabus, reflection on learning outcomes, reflection on teaching practices, assessment plan, course artifacts (assignments, handouts, etc.), assessment report, and final project reflection.

The Course Portfolio Working Group at Xavier has evolved over time to serve different institutional needs. We currently operate two separate groups – a traditional course portfolio group and a reading-oriented group that is associated with our Quality Enhancement Plan. And at one point in the past we offered both a "benchmark" portfolio group and an "inquiry" group. What follows here is a brief description of that history.

From 1997-2003, Xavier offered a "basic" course portfolio working group similar in many ways to what has been described above. This program was very successful-- so successful, in fact, that some faculty wished to repeat or continue the process. In response to these wishes in 2004-2005 we launched a new inquiry portfolio working group that we distinguished from the more basic benchmark group. The inquiry group was designed for graduates of the portfolio program and for more experienced faculty and its intent was to offer a portfolio experience that was more explicitly scholarly. Participants of the inquiry group were urged to view their work as an act of scholarship, that is, as a project that would culminate in a conference paper or publication. The inquiry portfolio group was effectively eliminated when Hurricane Katrina devastated New Orleans in August of 2005, though it could be revived in the future.

The most recent modification of the CPWG was the 2010-2011 introduction of a reading portfolio group, which developed in conjunction with the university-wide initiative called "Read Today, Lead Tomorrow." The RCPWG is the same as the regular portfolio group in most respects, except that its participants focus on learning outcomes and instructional practices that emphasize active and engaged reading.

Xavier's CPWG has drawn participants from across all disciplines, but the Pharmacy School has been among the most active supporters of the group. Pharmacy courses have a unique curricular context, because they are part of a doctorate program (students graduate with a Pharm.D. degree), and we are now planning to add a working group that is designed specifically for Pharmacy faculty.

The Course Portfolio Working Group as a Means to Improve Faculty Teaching

We have emphasized from the outset that course portfolios provide an effective vehicle for the assessment and improvement of teaching. Indeed, as outlined above, assessment is a major point of emphasis for the course portfolio working group at Xavier: it's where the fall semester finishes and where the spring semester begins. At the end of the year-long process, each faculty participant will have a data set that can be utilized to evaluate their teaching. This suggests the substantive richness and

broad application of the course portfolio, but in our experience the most beneficial aspect of the peer-driven working group experience is not the assessment of teaching *per se*, but the usefulness of assessment for the improvement of teaching. To put it in the language of assessment, it is not the summative assessment of teaching that is most valuable, but the more practical, formative assessment.

In terms of summative assessment, we encourage CPWG participants to consider viewing their portfolios (in whole or in part) as a documentation of the efforts they have made to improve their teaching or, even better, as a record of their teaching success. The portfolio then becomes something that may be submitted to the departmental chair or the rank and tenure committee for the purpose of a faculty evaluation or for decisions about tenure and promotion. This is no small matter, especially when we consider the limitations of the traditional end-of-semester student evaluations that are often the primary basis for judging faculty teaching.

Nonetheless, the course portfolio is even more valuable as an instrument of formative assessment, for it provides data that allow one to close the loop—that is, to make changes in order to improve one's teaching. If we think of teaching a class as a sort of educational intervention, that is, as an attempt to give students knowledge and skills that they formerly lacked, then assessment in this sense is the measurement of the value that has (or has not) been added. It provides instructors with information about the effectiveness of the instructional practices that they are using to accomplish the learning outcomes that define their courses. To become better teachers, we need more than a hunch or a feeling about our classes. We need specific information about what is and what is not working well; once armed with that knowledge, we can make changes to try and improve our teaching.

What kinds of assessment have worked best in our program? Ultimately this depends on the particulars of the course being taught and the person teaching it and we encourage participants in the CPWG to create their own individualized assessment plan. Some common assessment strategies include the following: customized end-of-semester student surveys; pre- and post-course comparative surveys; longitudinal analysis of student work; analysis of course work that targets specific content areas or skills; and electronic discussion boards and/or electronic portfolios that provide real-time glimpses of student work.

As this list suggests, assessment can take many forms and can provide different sorts of data. But to close the loop effectively, it is not the data that matters, but rather responding to the data in a way that will produce better teaching. The final phase of the CPWG at Xavier is devoted to precisely this issue; it is the subject of the last meeting of the year and the topic of the final portfolio reflection. What have participants learned and how should they respond to it? How can they bring that knowledge to bear on their courses in a way that will improve the courses? When participants pose those questions, and begin to answer them, they are closing the loop - and then the process can (at a higher level) begin all over again!

How to Get Started: Practical Advice for Institutions and Individuals

For Individuals

Embrace peer review. The course portfolio working group at Xavier is successful primarily because it provides a forum for faculty to interact with, learn from, and help each other. If your institution does not provide a similar forum, go out and find interested colleagues to work with. Your partner need not be in the same discipline; what's most important is to work with a colleague (or group of colleagues) that is

committed to the improvement of teaching. (See Ismail, Buskist, & Groccia, this volume, for a discussion of in-class peer review of teaching.)

Find good resources. There is an abundance of useful information available on course portfolios, including much that is available on the web. The reference list to this chapter details some useful resources. Perhaps the most helpful online resource is the Peer Review of Teaching Project at the University of Nebraska, Lincoln: www.courseportfolio.org. Go there first.

Emphasize process over product. Although it is nice to have the actual product in your hands at the end of a year's worth of hard work, what is most valuable is not the course portfolio itself, but the process of scholarly reflection on teaching. In this sense, the course portfolio is best understood as a continuing project rather than as a finished artifact and you will get more out of it if you approach it in this way.

For institutions

Cultivate administrative buy-in for formative assessment. For faculty to perceive that course portfolios are worth the time and effort, this activity must be an institutionally recognized and appreciated means of assessment. That is, portfolios must be valued by the administration. Cerbin (1994) argues that the weight that administrators place on typical summative assessments undermine activities that might actually improve teaching. If needed, educate your administration on the value of portfolios above and beyond end-of-term student evaluation data. Further, if possible, link your portfolio initiatives to institutional priorities (e.g., the strategic plan, the mission, accreditation requirements).

Provide incentives. Completing a meaningful, reflective portfolio takes time. If the institution views this activity as yielding valuable assessment results, then the initiative deserves resources either in the form of payment or course reduction. If budgetary restraints are an issue, perhaps it is possible to provide minimal funds for materials related to the course or simply a certificate of completion and recognition at a faculty meeting.

Develop leaders. In order for a systematic portfolio initiative to be sustained, there must be leadership in place. Train interested faculty on how to develop strong, meaningful portfolios and then allow them to facilitate working groups. Not only does this encourage more faculty to participate in this type of meaningful assessment, but it also builds faculty learning communities on campus.

Conclusion

Many benefits could be enumerated in support of the Course Portfolio Working Group at Xavier, starting with the scores of individual courses that have been re-invigorated, re-designed, and improved by dedicated faculty. And yet the best argument in support of this initiative is rooted not in individual courses or teachers, but in the campus culture that is created and fostered when faculty members work together to improve their teaching. The peer review nature of the course portfolio group builds real collegiality around a common purpose: teaching. It is a wonderful thing to see the physicist and the sociologist working together as allies for the common goal of teaching improvement. At Xavier this happens every semester in the Course Portfolio Working Group.

References

Boyer, E. (1990). *Scholarship reconsidered: Priorities of the professorate*. Princeton, NJ: Carnegie Foundation for the Advancement of Teaching.

Brookfield, S. D. (2006). *The Skillful Teacher.* San Francisco, CA. Jossey-Bass.

Cerbin, W. (1994). The course portfolio as a tool for continuous improvement of teaching and

learning. *Journal on Excellence in College Teaching, 5*(1), 95-105.
Cerbin, W. (1995). Connecting assessment of learning to improvement of teaching through the course portfolio. *Assessment Update*, *7*(1), 4-6. doi 10.1002/au.3650070103
Edgerton, R., Hutchings, P., & Quinlan, K. (1991). *The teaching portfolio: Capturing scholarship in teaching*. Washington, DC: American Association for Higher Education.
Gross Davis, B. (2009). *Tools for Teaching* (2nd ed.). San Francisco, CA: Jossey-Bass.
Hutchings, P. (1996). Portfolios: putting the pieces together. In P. Hutchings (Ed.), *Making teaching community property: A menu for peer collaboration and peer review* (pp. 49-60). Washington, DC: American Association for Higher Education.
Hutchings, P. (1998). Defining features and significant functions of the course portfolio. In P. Hutchings, (Ed.). *The course portfolio: How faculty can examine their teaching to advance practice and student learning.* (pp.13-18). Washington, DC: American Association for Higher Education.
Hutchings, P, & Shulman, L. (1999). The scholarship of teaching: New Elaborations, new developments. *Change*, *31*(5), 10-15. doi 10.1080/00091389909604218
Shulman, L. S. (1998). Course anatomy: The dissection & analysis of knowledge through teaching. In P. Hutchings, (Ed.), *The course portfolio: How faculty can examine their teaching to advance practice and student learning.* (pp. 5-12). Washington, DC: American Association for Higher Education.
Taras, M. (2005). Assessment – summative and formative – Some theoretical reflections. *British Journal of Educational Studies, 53*, 466–478. doi 10.1111/j.1467-8527.2005.00307.x

Contact Information

1. Paul Schafer, Associate Professor, Department of Philosophy, pschafer@xula.edu, 504-520-5406.
2. Elizabeth Yost Hammer, Director, Center for the Advancement of Teaching, eyhammer@xula.edu, 504-520-5141.
3. Jason Berntsen, Assistant Professor, Department of Philosophy, jberntse@xula.edu, 504-520-7624.

Peer Review of Teaching

Emad A. Ismail, William Buskist, and James E. Groccia
Auburn University

Peer review is a collaborative process by which teachers receive valuable feedback about their teaching performance from peers or colleagues. The concept of peer review received national attention following the 1994 collective initiative of 12 universities coordinated by the American Association of Higher Education (AAHE; Fernandez, & Yu, 2007). This initiative focused on providing formative feedback to and from "teams" of two instructors assessing course materials and teaching approaches in chemistry, mathematics, English, history, music, business, engineering, and nursing departments. Several evaluative strategies emerged from this work including the idea of "mutual mentoring" in which faculty (a) visit each other's classes to observe one another's teaching; (b) interview students; (c) review each other's teaching materials; (d) develop collections of departmental teaching materials for continuous course improvement; and (e) create colloquia in which job candidates present the design, pedagogical goals, and relevant assignments of a course they might teach (see also Hutchings, 1995).

Peer review can be used as formative method to offer instructors feedback about their teaching performance in a collegial atmosphere by providing information not typically included in students' evaluations (Victoria University of Wellington, 2004). Peer review can also be used as summative method for decision making about hiring faculty, granting tenure, making promotion decisions, deciding merit pay, and evaluating faculty teaching during post tenure review. The information, in the summative case, is for public inspection rather than for individual faculty use to improve teaching (Chism, 2007).

The Peer Review Process

Gosling (2002) identified three models of peer observation: The evaluation model (senior faculty observe other faculty), the developmental model (professional faculty developers observe faculty), and the peer review model (faculty observe each other). Our primary focus in this chapter will be on the last model because it is perhaps the most widely used version on college and university campuses and is most commonly misunderstood by college and university professors (e.g., Buskist, 2010; Chism, 2007; Perlman & McCann, 1998).

Many versions of the peer review model currently exist ranging from one faculty member (the teacher) simply asking another faculty member (the observer) to visit his or her classroom for observation and subsequent feedback, to a more thorough process involving five steps:

- the observer holding a preclassroom visitation meeting to discuss various aspects of teaching with the person to be observed,
- a classroom visitation to observe the teacher
- an opportunity for students to discuss and share their observations with each other and the observer regarding the teacher's teaching,
- a written report prepared by the observer for the teacher, and
- a postclassroom visitation meeting in which the observer provides both written and oral feedback to the teacher about what was observed during the classroom visit.

This multistep model of peer review allows for comprehensive analysis of one's teaching as well as the opportunity for extensive feedback on many aspects of what the observer witnessed during the classroom visit. In our view, this model is superior to all other models of peer review in terms of the

overall quality of support and teaching improvement feedback provided to faculty. Thus, the remainder of this paper will focus on the details of this model.

Step 1: Preclassroom Visitation Meeting

Before visiting the teacher's classroom, the observer holds a brief (usually less than 30 minutes) one-on-one meeting with the teacher to learn background information on her or his approach to teaching. The observer may or may not be acquainted with the teacher, but regardless, it is important that this meeting establish a cordial consultative relationship.

Scheduling of this meeting can be arranged by e-mail or telephone. Either way, it is important that the observer clarifies the meeting's purpose and asks for copies of (a) the course syllabus, (b) relevant course handouts or other materials, (c) the teacher's statement of teaching philosophy (d) the teacher's course portfolio (see Werder, 2000), and (e) if the teacher feels comfortable sharing them, recent student evaluations. The observer may request these materials in advance of the meeting in order to prepare for the preclassroom visitation meeting.

During this meeting, the observer asks for specific information about the teacher's approach to the class that will be observed. For example, observers may ask about the specific goals or student learning outcomes for the class session or the nature of the lesson or lecture plan with particular attention to how the teacher will help students achieve these outcomes. Gathering this information is especially helpful for observers who are reviewing instructors outside their academic specialty area and also helps observers understand and assess the extent to which the teacher and students actually accomplish the goals for the class session goals.

Observers also may ask teachers to describe any particular issues or problems that they are experiencing with the class. Having this information, in conjunction with already having examined the syllabus and student evaluations, helps observers attend to key aspects of the teacher's presentation and thus puts them in a better position to offer feedback regarding specific problems they observe.

If observers do not receive copies of the teacher's written materials ahead of the preclassroom visit, they can skim through these materials at the outset of the meeting. Observers may then ask to keep copies of these materials for later use in preparation for the postclassroom visitation feedback meeting. Observers often offer relevant positive comments on these materials (without any criticism) during the previsitation meeting; constructive feedback should be only offered during the postclassroom visitation meeting lest observers instill defensiveness in the teacher before the classroom visit.

At the end of the preclassroom visitation meeting, observers should establish the time and the location for both the class observation and the postclassroom visitation meeting. The postvisitation meeting should preferably be within 24 to 48 hours of the actual classroom observation. Finally, before leaving the preclassrroom visitation meeting, observers should emphasize the formative purpose of the observation and declare the confidentiality of all aspects of the peer review process.

Step 2: The Classroom Visitation—Observing the Teacher

The second step in the peer review process is the classroom visit. The observer should arrive at least 5 minutes before class, or arrange with the teacher to meet beforehand and walk to the classroom together—it is not appropriate to arrive to class late. The teacher should introduce the observer to students at the beginning of class and explain to them the nature of the observer's visit. Optimally, the observer should sit in middle rear of the classroom to get the best view of both the teacher's actions and the students' classroom behaviors. In addition to attending to particular issues that may have been

discussed during the preclassroom visitation meeting, the observer should pay attention to matters related to the teacher's delivery of content, physical presence, and social presence.

Delivery of content involves the manner in which the subject matter is presented during the class session and can be assessed through consideration of the following questions:

- Does the class begin and end on time?
- Does the teacher provide any sort of introduction to the subject matter or review material presented in the previous class session?
- Is the material presented at the appropriate level given the nature of the subject matter and the level of the class?
- Does the teacher communicate clearly with students—does the teacher explain jargon?
- Is the presentation logically organized—does the teacher employ useful transitions and examples to link or explain key points?
- Is the pace of the teacher's delivery about right—does it seem too fast or too slow?
- Does the teacher pose clear and interesting questions to the class?
- Are students' comments and questions repeated so that the entire class is able to hear them?
- If the teacher uses Power Point or a similar technology, is the font size legible and are the slides visually clear? Does the teacher avoid reading directly from the screen?
- Are appropriate demonstrations of the class material employed and, if so, are they related unambiguously to the subject matter?
- Does the teacher employ active learning techniques?

Physical presence centers on how teachers use their body language as a context to emphasize important points or to develop and maintain students' interest in the content. Aspects of the teacher's physical presence to which the observer should pay particular attention include:

- Eye contact with students
- Facial expression
- Movement about the room
- Posture
- Professional attire
- Hand gestures
- Voice—volume, inflection, and pace of speaking

Finally, social presence is the extent to which the teacher interacts appropriately with students. Factors that the observer should consider when assessing this category of variables include:

- Composure and confidence during the class session
- Reinforcement of student comments and questions with appropriate praise and language
- Level of engagement—does the teacher hold the students attention and interest?
- Respect for students
- Use of students' names

Observers, especially new observers, often find it helpful to bring an observation checklist with them to the classroom visit to serve as a reminder as to what teacher characteristics and behaviors and other aspects of the teaching situation they should be attending to during the visit. The checklist also provides space for the observer to jot down notes. Table 1 shows a sample of one such observation checklist

developed by the second author, who still uses it even though he has conducting peer reviews for over 20 years.

Table 1

A sample observation checklist used to help peer reviewers identify specific teaching behaviors and practices while observing a teacher.

Observation Checklist

Scale: 1 =Very Poor; needs serious substantial improvement
3= Good; needs a fair amount of improvement
5 = Excellent; needs little improvement

Content and Delivery	**1**	**2**	**3**	**4**	**5**	**N/A**	**Comments**
Appropriate use of time (begins/ends on time)							
Provides overview of topic/daily goals							
Appropriate level (depth & breadth)							
Clarity (prepared/explains jargon)							
Relevance (stays on topic)							
Knowledgeable & answers questions well							
Logical flow (organized & effective transitions)							
Pace of presentation/speaking							
Poses appropriate & clear questions							
Repeats students' questions/comments							
Uses of relevant examples							
PowerPoint (avoids reading off screen)							
PowerPoint (grammar & spelling)							
PowerPoint (font size & visual clarity)							
Use of demonstration/links to concepts							
Use of active learning techniques							
Handouts (useful in understanding topic)							
Provides conclusion/take home message							
Physical & Social Presence	**1**	**2**	**3**	**4**	**5**	**N/A**	**Comments**
Makes eye contact with students							
Facial expression							
Movement about room							
Posture							
Professional attire							
Uses appropriate hand gestures							
Voice—audible							
Voice—variation in inflection & tone							
Composure/confidence							
Reinforces student participation							
Has rapport with students							
Engaging (interesting and informative)							
Demonstrates enthusiasm							
Demonstrates respect for students							
Uses of student names							

Other Comments:

_____

_____

_____

Buskist (2000) has noted that teachers, especially new teachers, often commit particular errors in their teaching. Keeping an eye out for these mistakes may help observers spot and later offer corrective feedback for these mistakes. Based on his peer review/observation of new graduate student teachers and assistant professors, Buskist identified 10 common—but easily correctable—mistakes:

- Arriving late to class
- Starting the class "cold"/not providing an overview of the day's topic (launching abruptly into the subject matter without reviewing previous and related material)
- Returning tests/assignments at start of class (which puts some students in a "bad mood" and interferes with their paying attention for the remainder of the period)
- Reading directly from notes or PowerPoint slides
- Including too much information on slides
- Talking too fast
- Not using transitions between subtopics (not linking one topic to another in ways that help students see the connection between them)
- Not making eye contact with students (an important aspect of developing rapport with students)
- Not calling students by name (another important aspect of developing rapport with students)
- Not repeating student comments/questions or not rewarding student comments/questions with verbal acknowledgement or praise.

These qualities and behaviors are present in any teaching situation in manifold combinations and permutations—no one teacher will be exactly like another, although they both may be excellent (or poor) teachers. Thus observers should keep an open mind as how these qualities and behaviors may play out in any given classroom session and should not enter the classroom observation with a preconceived notion of what they expect to observe.

Step 3: The Class Visitation—Talking to the Students

A limitation to peer review is that it often provides only a single snapshot of the teacher's instructional skills; it cannot provide information about what the learning experience has been for students up to that point in time. To offset this problem, it is often helpful for the observer to talk to the students about the teacher's instructional approach and what the students' learning experience has been like so far in class. Such information can be useful in painting a larger picture of the teacher's strengths and needed improvement, which places the observer in a much more informed position for offering useful feedback during the postclassroom meeting.

The most common method for observing teachers and talking to students is to arrange for both to occur on the same day. Thus, in a typical class period, actual observation lasts for about 25 minutes after which the teacher leaves the room and the observer begins a conversation with students for the remainder of the class period. For longer class periods than the standard 50-minute variety, the observer and the teacher may discuss an arrangement for combining the observation and the discussion with the students so as to maximize the effective use of the class period.

Once the teacher leaves, the observer moves to the front of the room and reintroduces himself or herself to the students and lays out the "ground rules" for the remainder of the period. The observer stresses that this meeting is not a complaint or gripe session, but rather an opportunity to comment on positive aspects of the teacher and the class and to offer constructive comments on how the course might be improved to enhance students' learning of the subject matter. The observer also notes that the content of the discussion will be shared only with the teacher and that their identities will not be revealed.

In general, the discussion with the students [also known as Small Group Instructional Diagnosis (SGID) or Feedback (SGIF)] follows five steps:

1. The observer divides the students into small groups (4-6 students for small classes; 7-10 students for large classes).
2. The observer either writes on the board or hands out a piece of paper to each group with the following three questions written on it:
    a. What is going well in this class so far? Or, what do you like about your instructor's teaching? (This item reveals what the teacher should keep doing.)
    b. What aspects of your teacher's teaching or the course need improvement? (This item lets the teacher know what he or she might change or start doing.)
    c. What other comments do you have? (This item relates to aspects of the course not covered by the two previous questions such as the quality of the learning environment.)
3. The observer asks one student in each group to take notes regarding its discussion of the three questions. The observer gives each group 6-8 minutes to discuss the three questions.
4. The observer then asks each group to stop and initiates a whole-class discussion over each question. The observer asks one student to serve as scribe for this discussion by taking notes for the remainder of the class period. Each question is thus discussed in order, with the observer repeating students' comments aloud, and if necessary asking clarifying questions. The observer should not offer any evaluative comments or show emotional reactions to students' commentary.
5. Once students have had a chance to share their responses to each question, the observer collects all students' notes (the notes taken during the small group discussion as well as those taken during the larger class discussion), thanks students for their time and comments, and dismisses the class. The entire discussion with the students generally takes no longer than 20 minutes. Nonetheless, the observer should keep an eye on the clock so as make sure that the discussion is completed on time.

Step 4: Preparing Written Report

Shortly after the classroom visitation and in preparing for the postclassroom visitation meeting, the observer combines the students' commentary with his or her observations and notes taken while the teacher was actually leading the class and prepares a written report. In writing the report, the observer should (a) note, in detail, the teacher's strengths; (b) outline the teacher's key areas in need of improvement (using as much positive language as possible); and (c) offer specific suggestions for addressing each area in need of improvement. Along these lines, a good rule of thumb is to only comment and offer suggestions for the two to three areas, which if addressed, would result in the most immediate and tangible improvement in student learning and student enjoyment of the course (addressing every area in need of improvement, especially if there are many, might be overwhelming for the teacher and actually discourage change).

Step 5: Postclassroom Visitation Meeting

The final step of the peer review process is the observer's postclassroom meeting with the teacher. This final step involves one-on-one discussion about what went well during the class and what might be improved in the teacher's approach to instruction. It is based entirely on the observer's written report.

This meeting usually lasts for less than one hour and should transpire as a conversation between two people who respect each other and care for teaching. The meeting should not simply be a word-for-word reading of the observer's written report. The observer may start by asking, "What do you think went well during the class I observed?" This sort of beginning to the conversation helps the teacher to focus on his or her strengths on which the observer can build. This beginning can also reinforce what the teacher says and add to the description of positive qualities and behaviors noted in the written report. The observer should also add any relevant points brought up in the discussion with the students.

Once both parties have discussed the teacher's strengths, the observer transitions to a discussion of areas of improvement by noting, for example, "Although you have several important strengths as a teacher, students have identified a few areas that might be improved." Or, "As is true for all of us, we have both strong points as well as a few areas in which we might improve—let's talk about a couple of things that the students suggested that might help improve your teaching." It can be helpful, before the observer shares his/her interpretations of what students have suggested, to ask the teacher to predict what students said in this regard. This tactic helps the teacher to become more personally involved and invested in the review process. We have found that teachers can often anticipate what suggestions students offered, even while not feeling capable or empowered to make such changes. During this part of the discussion, the observer should pay careful attention to both the teacher's verbal language and body language, and make any adjustments necessary if the teacher appears to be uncomfortable with this part of the meeting.

The observer should avoid using language that is strong, harshly critical, or authoritarian. It is important that the observer be as gentle, but frank, as possible with the teacher in order to have maximum impact improving the teacher's instruction. Otherwise, the teacher may "tune out" the feedback, and the meeting might devolve into an unpleasant experience for both the parties.

During the discussion of areas in need of improvement, it also is often helpful for the observer to share any personal insights or experiences that might provide clarity or exemplars as to how to address these areas. We have found that using such lead-in language as "I faced the same issue in my teaching and found that _____ helped me improve in this area," or "I had a discussion with a colleague the other day about a similar issue she (or he) faced, and she did _____ to resolve the issue." Using this kind of approach lets teachers know that others also experience similar issues and that none of us are alone in discovering that some areas of our teaching need improvement.

As the conversation winds down, the observer should request that the teacher let the students know about this meeting, what was discussed during it, as well as any changes that can or cannot be made in the class based on student input to the peer-review process. In our experience, we have found that students like having a voice in shaping the nature of their class. They like knowing that their teacher is concerned about the quality of the classroom atmosphere and instruction, and they genuinely appreciate any changes made based on their comments and advice.

At the conclusion of the meeting, the observer should provide the teacher a hard copy of the written report. The report should not be given to the teacher before the meeting because, after reviewing it, the

teacher may focus only on the areas of improvement (i.e., the "negative points") and become defensive at the outset of the meeting. For confidentiality reasons, the observer should not keep copies of the report (neither hard nor electronic) after the meeting. The peer reviewer may however, keep a confidential copy of the report for a given, mutually agreed upon period of time (e.g. one semester) in case the teacher misplaces his/her copy.

Four Common Questions about Peer Review

We have outlined the key steps in one method of conducting peer review. The remainder of this chapter will address four questions that often arise in discussion of any kind of peer review that takes place at the college and university level.

Who should conduct the peer review?

A fair and effective peer review requires many skills and personal attributes in faculty observers (e.g., Buskist, 2010; Chism, 2007; Perlman & McCann, 1998). Peer reviewers should be enthusiastic about teaching and about students, knowledgeable about pedagogy and what constitutes good teaching, and be nonjudgmental. They should also be careful observer of human behavior—both verbal and non-verbal varieties— a reflective listener, and possess good social skills. Brent and Felder (2004) proposed that peer reviewers should be tenured colleagues or faculty or non-tenure-track faculty with primarily teaching and advising responsibilities who are competent and broad-minded teachers when it comes to considering what is and what is not effective pedagogy.

Most peer review focuses on helping teachers develop more effective teaching skills—the emphasis is not necessarily or primarily on the teacher's knowledge of the subject matter. Thus, oftentimes, peer review is conducted with peers from different disciplines. An added benefit of having faculty from a different discipline serve as peer reviewers is that it sidesteps the almost unavoidable social pressures for the observer to be "nice," or at least less than frank, when critiquing a colleagues' teaching, lest feelings get hurt or a relationship become stressed (Groccia, 1998).

What should peers review?

The answer to this question depends on who is doing the review. If the review is being conducted by a non-disciplinary peer, the focus is almost always placed on elements of the teaching process that we alluded to earlier: preparation and organization for daily classroom teaching (e.g., course syllabus, student learning objectives, lecture plan), teaching method and delivery of the content, formative and summative assessment techniques of student learning, communication skills, physical presence (e.g., body language, voice, facial expressions, physical mannerisms), social presence, classroom management techniques, the teacher's personal characteristics (e.g., enthusiasm, sense of humor), and use of appropriate technologies. In the most thorough of peer reviews, the peer observer will also review the teacher's statement of teaching philosophy and a course portfolio, if one has been created for the course (Bowser, 1997).

However, if the review is being conducted by a disciplinary peer, then the focus is often on the teacher's subject matter knowledge and delivery of content. For example, a disciplinary peer would pay particular attention to the appropriateness of course objectives, currency and accuracy of the subject matter, knowledge of what aspects of the subject matter should be taught in any particular course, application of the most appropriate methodology for teaching this specific content, and appropriateness of student assessment tools given the subject matter and student learning objectives for the course (Chism, 2007; Cohen & McKeachie,1980; Fernandez & Yu., 2007; Keeley, 2012).

What are the benefits of peer review?

Knowing the subject matter well does not mean one can teach it well. Teachers achieve excellence by mastering the subject matter and mastering methods used to communicate that content clearly and unambiguously to students (Buskist, Sikorski, Buckley, & Saville, 2002). Based on our combined 65 years of working in higher education, we feel that peer observation, especially when used for formative purposes, can contribute strongly to enhancing the quality of one's teaching. For example, research has shown that in-class observation or videotaping graduate teaching assistants while teaching followed afterwards by consultation with their teaching advisors is effective in improving teaching (e.g., Abbott, Wulff, & Szego, 1989; Lewis, 1997; Shore, 2012).

If conducted by mid-semester, peer review, particularly when combined with feedback from discussions with students, provides teachers the opportunity to adjust their approach during the same semester and before receiving end-of-semester students' evaluations. As we have discovered at Auburn University, an added benefit of early peer review is that making changes based on the feedback from peer observation and student commentary, improves the teacher's scores on the end-of-the-semester evaluation (see Wilson & Ryan, this volume).

Peer review has reciprocal benefits for both parties. Of course, teachers benefit from the constructive feedback that they receive from observers about their approach. Likewise, observers often benefit by learning new teaching techniques and by observing new approaches to issues and problems observed during the peer review, and by learning from student discussions about the class. In a sense, peer reviewers get an opportunity to enhance their own teaching through self-assessment stimulated by watching how others teach and gathering new ideas for their own critical thinking about teaching (Bovill, 2010; Gosling, 2005). In a similar vein, peer review often fosters collegial discussion of teaching and the dissemination of good teaching practices among faculty, which allows for potential cross-fertilization of ideas and approaches to teaching (Gosling, 2005).

What are the putative limitations of peer review?

Despite the obvious utility of peer review in improving the quality of teaching and student learning, many academics have raised concerns regarding the practice of peer review. These concerns range from faculty discomfort in being observed to the belief that peer review infringes on academic freedom (Chism, 2007). We briefly address the more common of these concerns below.

Teaching is an unimportant faculty activity. It is commonly assumed among many individuals across the academy that teaching is less important than research, particularly when tenure, promotion, and merit salary increases are concerned. Thus, as the argument goes, faculty resources should be aimed at supporting research more so than teaching (Berk, 2005; Keig & Waggoner, 1995; Lomas & Kinchin, 2006). This mindset is generally more prevalent at large research-centered universities than it is at smaller liberal arts and community colleges.

Our counterargument to this claim is that regardless of where it occurs, undergraduate teaching is a critically important responsibility of the academy. All academics who are teachers have a responsibility to offer their students the best possible learning experience in their classes. Indeed, we and others consider it unethical to do otherwise (e.g., Hill & Zinsmeister, 2012).

Peer review is only summative in nature. A common misconception is that the purpose of peer review is to provide an evaluation of faculty teaching to be used in making summative decisions about tenure, promotion, merit, and so on. Although it is true that peer review can and has been used for this

purpose, it can also be used for formative purposes. Indeed, we argue that the primary benefit of peer review is the feedback given to faculty to help them improve the quality of their teaching and learning experiences they provide for their students. At Auburn University, the Biggio Center for the Enhancement of Teaching and Learning conducts dozens of peer reviews across campus each semester, and all of them focus wholly on formative peer review. Likewise, teaching preparation in Auburn University's Psychology Department focuses solely on formative peer review (provided by faculty members). None of the information provided during the peer review process is shared with administrators.

Good teaching can be faked. In our workshops on peer review, some faculty have argued that any teacher can pretend to be a good teacher for one class period (see also Brent & Felder, 2004). This claim holds that what a peer reviewer may see during the classroom visitation is not representative of what the teacher is truly like on a daily basis. We agree that most teachers who are observed will try their best to teach well during the visit. We have three responses to this issue.

First, teaching well is a difficult and complicated endeavor (Groccia, 2012; Svinicki & McKeachie, 2011) and, because classroom observation generally ranges from between 25 and 50 minutes, it is difficult for many, if not most, faculty to be "perfect" in every aspect of their teaching for that long. To be sure, they may do many things well, but of course, they also likely will do some things not so well. In particular, distractive mannerisms, including both verbal and body language, may manifest themselves, as will issues related to organization, clarity, and their choice of examples. In our combined experience of peer reviewing, we have yet to run across the perfect lecture or class presentation, faked or genuine.

Second, we believe just the very fact that they will be observed motivates faculty to improve their teaching, with positive outcomes for both faculty and students. By trying to create an outstanding lecture or classroom presentation on the day of peer review, faculty may discover new and improved ways of teaching and find that they enjoy the outcomes of doing so. Future class sessions may be improved as a result.

Finally, meeting with the students following the classroom observation can clear up any discrepancies between what the peer reviewer observed and what students have typically experienced with that particular teacher so far in the semester. During our classroom visitation conversations with students, we will often ask them if their teacher's approach during the observation was representative of what they have experienced thus far in the course. Seldom do students reply that the teacher's behavior on the day of the observation was atypical.

Classroom performance is not all there is to teaching. As we noted at the outset of this chapter, most faculty think of peer review as focused wholly on classroom performance—and it is true that peer review takes this form on many occasions. However, the format for peer review we have outlined in this chapter is much more expansive than that. It includes a preclassroom observation meeting with the teacher to review course syllabi, course evaluations, statements of teaching philosophy, and other documents that help the peer reviewer capture the "bigger picture" of the teacher's approach to teaching. It also includes a classroom meeting with students to gather their thoughts on their teacher's total approach to the course they are taking. Thus, peer review can focus exclusively on a teacher's classroom performance or it can include, in addition observing to classroom performance, collecting additional data and information from both teachers and students.

Most faculty are not experts in peer review. We could not agree more. The fact that peer reviewing entails more than just being a subject matter expert is sufficient to support this claim—peer reviewing demands having considerable understanding of effective pedagogy and classroom management practices. Most faculty are not experts in conducting peer review and those who conduct peer review should receive sufficient training in teaching and classroom management practices to qualify them to practice competent peer review. Until college and universities require this kind of training, the quality of peer review and its potential impact on students' learning experiences will remain limited.

Conclusion

We have outlined a five-step approach to peer review: (a) a preclassroom visitation meeting with the teacher to learn about his/her approach to teaching in general and the class to be observed in particular, (b) the classroom observation in which the peer reviewer observes the teacher conducting class, (c) a conversation with his/her students about their learning experiences in the class; (d) the observer's preparation of a written report, and (e) the postclassroom observation meeting between the peer reviewer and the teacher to discuss the latter's strengths and areas in need of improvement based on the observation and student discussion. Although academics have raised several concerns about the limitations of peer review, we suggest that in some cases, these concerns are false, and in other cases, simple solutions exist to correct the concern. However, one serious limitation of peer review is that many faculty asked to perform peer review may not be "experts" in what constitutes good teaching. Thus, these individuals may not be in an optimal, or even sufficient, position to offer clear advice on how the teachers they observe might improve their instruction. We call on all college and universities to offer their faculty opportunities to become fully trained in college and university pedagogy so they can conduct high quality peer reviews of teaching.

References

Abbott, R. D., Wulff, D. H., & Szego, C. K. (1989). Review of research on TA training. In J. D. Nyquist, R. D. Abbott, & D. H. Wulff (Eds.), *Teaching assistant training in the 1990s.* New *Directions for Teaching and Learning*, No. 39 (pp. 111–124). San Francisco, CA: Jossey-Bass.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17,* 48-62. Retrieved from http://www.isetl.org/ijtlhe/pdf/IJTLHE8.pdf

Bovill, C. (2010). *Peer observation of teaching guidelines*. Glasgow, Scotland: Learning and Teaching Centre, University of Glasgow. Retrieved from http://wiki.ubc.ca/images/0/0b/Peer_Observation_of_Teaching_Guidelines_-_Bovill_C.pdf

Bowser, B. (1997). *University of North Carolina intercampus dialogues on peer review of teaching: Results and recommendations*. Retrieved from http://www.uncg.edu/tlc/downloads/unc_dialogue.pdf

Brent, R., & Felder, R. M. (2004). *A protocol for peer review of teaching.* Session 3530. Proceedings of 2004 Engineering Education Annual Conference & Exposition (ASEE).Salt Lake City, Utah. Retrieved from http://uwf.edu/CAS/partners/documents/Peer_Review_Protocol_ASEE04.pdf

Buskist, W. (2000). Common mistakes made by graduate teaching assistants and suggestions for correcting them. *Teaching of Psychology*, *27,* 280-282. doi: 10.1207/S15328023TOP2702_13

Buskist, W. (2010). *Peer consultation manual*. Riyadh, Saudi Arabia: King Saud University, Deanship of Skills Development.

Buskist, W., Sikorski, J., Buckley, T., & Saville, B. K. (2002). Elements of master teaching. In S. F. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 27-39). Mahwah, NJ: Erlbaum.

Chism, N. V. N. (2007). *Peer review of teaching. A sourcebook* (2nd ed.). San Francisco, CA: Jossey Bass.
Cohen, P. A., & McKeachie, W. J. (1980). The role of colleagues in the evaluation of college teaching. *Improving College and University Teaching*, *28*, 147-154.
Fernandez, C. E., & Yu, J. (2007). Peer review of teaching. *Journal of Chiropractic Education. 21*, 154–161.
Gosling, D. (2002). *Models of peer review of teaching*. London, England: Learning and Teaching Support Network.
Gosling, D. (2005). *Peer observation of teaching* (Paper 118). London, England: Staff and Educational Development Association.
Groccia, J. E. (1998). *Overview of peer review of teaching*. Biggio Center for the Enhancement of Teaching and Learning, Auburn University. Retrieved from http://www.auburn.edu/academic/other/biggio/services/overviewpeerevaluation2009.pdf
Groccia, J. E. (2012). A model for understanding university teaching and learning. In J. E. Groccia, M. A. T. Alsudairi & W. Buskist (Eds.), *Handbook of college and university teaching: A global perspective* (pp. 2-13). Thousand Oaks, CA: Sage.
Hill, G. W., IV, & Zinsmeister, D. D. (2012). Becoming an ethical teacher. In W. Buskist & V. A. Benassi (Eds.), *Effective college and university teaching: Strategies and tactics for the new professoriate* (pp. 125-137). Thousand Oaks, CA: Sage.
Hutchings, P. (Ed). (1995). *From idea to prototype: The peer review of teaching, a project workbook.* Sterling, VA: Stylus.
Keeley, J. (2012). Course and instructor evaluation. In W. Buskist & V. A. Benassi (Eds.), *Effective college and university teaching. Strategies and tactics for the new professoriate* (pp. 173-180). Thousand Oaks, CA: Sage.
Keig, L., & Waggoner, M. (1995). Peer review of teaching: Improving college instruction through formative assessment. *Journal on Excellence in College Teaching, 6,* 51-83.
Lewis, K. G. (1997). *Training focused on postgraduate teaching assistants: The North American model.* Retrieved from www.ntlf.com/html/lib/bib/backup/lewis.htm
Lomas, L., & Kinchin, I. (2006). Developing a peer observation program with university teachers. *International Journal of Teaching and Learning in Higher Education, 18*, 204-214. Retrieved from http://www.isetl.org/ijtlhe
Perlman, B., & McCann, L. I. (1998). *Peer review of teaching: An overview.* Society for the Teaching of Psychology (APA Division 2), Office of Teaching Resources in Psychology (OTRP), Retrieved from teachpsych.org/otrp/resources/perlman98.pdf.
Shore, C. M. (2012). Assessing the effectiveness of GTA preparatory activities and programs. In W. Buskist & V. A. Benassi (Eds.), *Effective college and university teaching. Strategies and tactics for the new professoriate* (pp. 181-187). Thousand Oaks, CA: Sage.
Svinicki, M., & McKeachie, W. J. (2011). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (13th ed.). Belmont, CA: Cengage.
Victoria University of Wellington. (2004). *Improving teaching and learning. Peer observation of teaching*. Wellington, New Zealand: University Teaching Development Centre, Victoria University of Wellington, Retrieved from http://www.utdc.vuw.ac.nz/resources/guidelines/PeerObservation.pdf
Werder, C. (2000). *How to prepare a course portfolio*. Western Washington University, Center for Instructional Innovation and Assessment. Retrieved from http://pandora.cii.wwu.edu/cii/resources/portfolios/preparation.asp

Contact Information

1. Emad A. Ismail is a graduate teaching assistant at the Biggio Center for the Enhancement of Teaching and Learning and may be contacted at eam0028@tigermail.auburn.edu

2. **William Buskist** is the Distinguished Professor in the Teaching of Psychology and a Faculty Fellow of the Biggio and Center for the Enhancement of Teaching and Learning and may be contacted at buskiwf@auburn.edu

3. **James E. Groccia** is Director of the Biggio Center for the Enhancement of Teaching and Learning and Associate Professor of Higher Education at Auburn University in Auburn, Alabama and can be contacted at groccje@auburn.edu