Student Evaluations of Teaching: Perceived Merits and Disadvantages, and Suggestions for Improving the Assessment Method

James E. Miller, Ph.D.

Harding University

Abstract

The use of student evaluations of teaching (SET) to assess instructors' effectiveness is one of the most common and controversial practices in higher education. While a number of researchers have concluded that SET's are a valid, reliable, and worthwhile means of assessment (Wachtel, 2005; Koon & Murray, 1995; Centra, 1993), detractors contend that the method is too narrow in focus and open to bias. This paper explores the history of SET's, analyzes perceived benefits and disadvantages of the assessment method, and offers suggestions for improving this popular means of measuring teacher effectiveness.

Introduction

Quality assurance in the higher learning has been an important and often debated topic of discussion for decades. Much has been written about issues of accountability in the academy, including institutional and program accreditation, financial aid opportunities for students, and public funding of higher education, just to name three. Certainly, the areas within the higher learning for which the public and government have demanded accountability comprise a long list. The list of proposed methods for evaluating the effectiveness of higher educational institutions and their programs is, presumably, even longer.

One of the areas of quality assurance that researchers and authors have addressed consistently over the last 40 years is teaching in the academy. At a time when an increasing number of universities seem to value research productivity more than teaching effectiveness, it is no wonder this subject continues to receive significant attention in higher learning circles. In this paper, the author explores the predominant means for assessing instructor performance in the college and university classroom: student evaluations of teaching (SET's).

It must be stated at the outset that shedding new light on this controversial subject is not the primary purpose of this report. Instead, the following pages serve to describe and critically analyze the use of SET's in the academy based on the literature that exists in the field. And there is no shortage of literature that addresses student evaluations of teaching; more than 2,000 published studies exist on the subject, according to Murray (2005). Moreover, Cashin (1988) estimated that SET's have been studied more than all other forms of college teaching evaluation methods combined.

As a result, the author of this present paper intends to (1) describe the history, purpose, and characteristics of this evaluative practice in higher education, (2) discuss the controversy surrounding the use of SET's, including perceived merits and disadvantages of the assessment method, and (3) offer broad suggestions for improving this popular means of measuring instructor effectiveness. The analysis found in the following pages provides a foundation for the construction of novel ideas that may improve the methods for assessing classroom teaching in the academy. That lofty possibility makes this present descriptive and analytical work beneficial to all stakeholders of higher education.

History and Purpose of Student Evaluations

Although SET procedures were introduced in a few major universities in the 1920s (Marsh, 1987), student evaluation of teaching became a regular assessment tool in most North American institutions of higher education in the late 1960s and early 1970s. Today, more than 90% of U.S. colleges and universities use some sort of student evaluation mechanism to assess teaching (Murray, 2005).

The desire to implement a measurement of teaching effectiveness based on student feedback is understandable and commendable. After all, students are one of the consumer groups interested in the product of a college or university education; therefore, their opinions are a vital source of information concerning the quality of instruction at institutions of higher education (Wright, 2006).

However, students are not the only constituent group interested in the evaluation of classroom teaching. Faculty and administrators – not to mention important outside groups such as alumni, donors, legislators, and taxpayers – have motives for desiring a sound method for assessing instructor effectiveness. To illustrate this point, Murray

(2005) recalls that when the University of Western Ontario began using SET's in the late 1960s, it did so with the support of three parties: (1) students who wanted a say in matters of teaching, (2) administrators who were concerned with accountability and positive public relations, and (3) junior faculty who wanted their salaries, promotions, and tenure decisions to depend on something more than their research alone.

Not surprisingly, even in the early days of implementation, many senior faculty members who were more interested in boasting strong research records than teaching success met SET's with some resistance. As more and more institutions started using student evaluations as a basis for tenure and promotion decisions, criticism of the reliability, validity, and integrity of SET's increased.

Nonetheless, despite being troubled by many aspects of student evaluations, most college instructors (about 70%) agree on the need for student input into the assessment of their teaching (Obenchain, Abernathy, & Wiest, 2001). It is interesting to note, however, that Avi-Itzhak and Kremer (1986) discovered senior and tenured faculty were most opposed to the use of student evaluations for summative purposes. They logically concluded that this is a result of senior professors' spending more time on research and less time on teaching, which makes them less student-oriented than their junior faculty counterparts.

Of course, even if few professors – junior or senior – favored student evaluations, they would find it difficult to avoid such assessments. As Lewis and Benson (1998, p. 99) posit: "You cannot flee from the evaluation of teaching and you should not try to do so. We suggest you embrace the process and learn from it." Undoubtedly, general support for SET's among higher learning stakeholders is tempered with the understanding that

student evaluations actually serve as an alternative to the ideal means of assessing teaching effectiveness – direct measurement of student learning (Murray, 2005).

Because assessing student learning is "fraught with technical difficulties," student evaluations of teaching attempt to do the next best thing by evaluating teacher or course characteristics that are: (1) believed to contribute to student learning, based on evidence or logical argument, (2) observable by students, (3) widely applicable, and thus can be used in many different courses, and (4) under the control of the instructor, and thus are justifiable for use in faculty personnel decisions on salary, promotion, and tenure (Murray, 2005, p. 2).

As Murray (2005) rightly concedes, however, applying these criteria severely limits the range of teacher and course characteristics that can be included on a teacher rating form, and, as a result, imposes significant limitations on the integrity of student evaluations. Perhaps the simplest and most optimistic purpose statement for the teaching evaluation process is that it exists to improve teaching. As Gallagher (2000, p. 141) states, "In the absence of feedback, instructors would have to rely exclusively on their own inferences about the quality of their teaching."

Of course, just because students complete evaluations of teachers does not release instructors from the responsibility of appraising their own teaching effectiveness while a class is underway. Accordingly, Gallagher (2000) advises teachers to ask themselves questions such as: (1) To what extent do students participate in class by asking questions and offering insights? (2) To what extent are students attending class? (3) How well are students learning the material as measure by exam scores?

Characteristics of Student Evaluations

Of course, the usefulness of SET's to instructors hoping to improve their teaching depends on the content and coverage of the items included on the evaluative instrument. These instruments most often measure effectiveness solely by quantitative standards (Wallace & Wallace, 1998). A typical teacher rating form allows students to score statements on a five-point scale that address the instructor's clarity of expression, enthusiasm, availability, and fairness on exams (Murray, 2005).

However, as Marsh and Roche (1997, p. 1187) assert, "Poorly worded or inappropriate items will not provide useful information, whereas scores averaged across an ill-defined assortment of items offer no basis for knowing what is being measured." Moreover, valid measurement demands a "continual interplay between theory, research, and practice," according to Marsh and Roche (1997, p. 1187), who argue that evaluations of teachers should reflect the complex activity of teaching, which consists of multiple dimensions.

Accordingly, Marsh and Roche (1997) posit that the Students' Evaluation of Educational Quality (SEEQ) instrument most effectively assesses an instructor's multidimensionality. The SEEQ, whose creation was based on reviews of current instruments and interviews with students and faculty, evaluates teachers and courses on nine factors: learning/value, instructor enthusiasm, organization/clarity, group interaction, individual rapport, breadth of coverage, examinations/grading, assignments/readings, and workload/difficulty.

Although, as mentioned earlier, most SET's are quantitative in nature, many institutions provide students opportunities to write in great detail comments relating to an instructor's effectiveness. While this attempt at qualitatively assessing teacher

performance is applauded, rarely do administrators take these forms into consideration. For example, the author taught at an institution where the quantitative ratings were scored, compared across departmental and institutional means, and kept on file at the institutional and departmental levels. While some administrators may have taken note of the qualitative forms (it certainly varied from college to college and department to department), individual instructors typically paid more attention to the written comments than did their superiors.

In the end, the SET's had little effect at the institutional level. The university simply filed the evaluations in the on-campus office responsible for managing SET's and then sent individual instructor's results to the appropriate academic departments. In other words, teacher evaluations did not have any bite at the institutional level. Sure, some departments chose to reward teachers who received consistently high SET scores, or provide assistance to instructors who received consistently low scores, but other departments simply filed the student evaluations without consequence.

When SET's were first implemented at the aforementioned institution, the Faculty Senate voted not to consider student evaluations in tenure and promotion decisions. While this remains the official policy, some departments do take SET's into consideration, if not unintentionally. At other colleges and universities, however, salary, promotion, and tenure decisions within individual academic departments are made with official consideration given to student evaluations of teaching. However, Murray (2005) and Read, Rama, and Raghunandan (2001) found that while teaching does have an impact on salary, promotion, and tenure decisions, it is small compared to the impact of

research productivity. In fact, student evaluation of teaching accounts for only about 10% of the variance in personnel decisions among faculty (Murray, 2005).

While some institutions choose to limit the ramifications of SET's, others are considering ways to promote the results of student evaluations so that anyone can view them. In an age where websites such as ratemyprofessor.com have made it easy for students to access and participate in professorial and course reviews that have no ties to universities and, as some claim, lack quality control, a few institutions are toying with the idea of publishing their official SET results (Epstein, 2006).

At Northwestern University, for example, the faculty senate discussed several years ago the merits of publishing online all instructors' SET scores and student comments. At the time, Northwestern professors could give consent to the university to publish basic data taken from student evaluations, and only about 5% had asked that their evaluations not be posted online (Epstein, 2006). Proponents of the idea argue that times have changed since the Northwestern faculty first agreed to a student evaluation process – an agreement that allowed the faculty to keep the results under lock and key.

Today, says Stephen Fischer, associate provost for undergraduate education at Northwestern and advocate of publishing student evaluations of all instructors, student evaluations "have become commonplace on campus. Students look at them, and so do members of the administration and faculty members" (Epstein, 2006, p. 1). However, opponents say publishing all results – especially students' written comments regarding a teacher's performance in the classroom – may do more harm than good. As Fischer stated in Epstein's (2006, p. 1) article:

> One communications faculty member said there are studies that say negative comments linger longer than positive comments, and that a few bad comments could color the perception of that teacher disproportionately. If it's numerical, [a small number of scores] would be dissipated.

Certainly, the contention over whether to publish student evaluation results in their entirety for anyone to view is just one factor in the encompassing debate over SET's, which this paper will now address.

Arguing the Validity and Reliability of SET's

As mentioned earlier, an enormous amount of literature addresses student evaluations of teaching. Moreover, in the thousands of studies that researchers have conducted, the predominant conclusion is this: SET's are a valid, reliable, and worthwhile means of evaluating teaching (Centra, 1993; Cohen, 1981; Koon & Murray, 1995; Marsh, 1987; Marsh & Dunkin, 1992; McKeachie, 1990; Ramsden, 1991; Seldin, 1993; Wachtel, 2005). In fact, Marsh (1987) posits that student evaluations are the only measure of teaching effectiveness whose validity has been thoroughly and rigorously determined.

Murray (2005) summarizes the literature that investigates the integrity of SET's by reporting that student ratings are sufficiently reliable, in that ratings of instructors are reasonably consistent across courses, sections, years, rating forms, and groups of raters. Additionally, Murray (2005) attests to the validity of student evaluations, in that they generally agree with evaluations made by others, such as colleagues and alumni, and are relatively free of bias. However, he argues the strongest defense of the validity and reliability of student evaluations comes in two specific types of studies.

The first includes classroom observations studies in which trained observers visit classes to record teaching behaviors. Once the observers have reported on the instructors' behaviors, an attempt is made to predict student assessments of teaching from those outside reports. Murray (2005) cites data that show student ratings are closely related to and highly predictable from specific classroom behaviors of the instructor, thus reinforcing the validity of SET's.

Second, Murray (2005) discusses studies that investigate multi-section courses. A multi-section course as defined in these studies has different instructors but a common syllabus, curriculum, final exam, and final exam grading system. Accordingly, it is assumed that differences in section mean scores on the final exam reflect differences in amount learned by students in the specific classes, rather than differences in instructor grading practices or curriculum choices throughout the term.

The question driving these multi-section course studies is, "Do instructors who receive high ratings from students actually teach their students more effectively so that they perform better on the common final exam?" According to Murray (2005), the answer is "Yes." The data show that students taught by highly rated teachers tend to learn the subject matter better than those taught by lower rated teachers. In other words, Murray (2005, p. 3) contends, "student ratings validly reflect differences in actual teaching effectiveness, rather than extraneous variables." Indeed, many researchers have reached similar conclusions (Aleamoni & Hexner, 1980; Centra, 1977; Cohen, 1981; McKeachie, 1990).

Along with believing student evaluations of teaching are reliable and valid, Murray (2005) boldly asserts that they have contributed to improved teaching in the

higher learning that is of better quality than it was 30 or 40 years ago. Apparently, he is not the only one to make this determination: 73.4% of faculty agree that student evaluations of teaching provide useful feedback for instructors, and 68.8% agree that SET's have improved teaching (Murray, 2005). Additionally, many other scholars have advanced the notion that feedback from student ratings can help improve instruction (Cohen, 1980; Menges, 1991; Overall & Marsh, 1979).

Similarly, Aleamoni (1981) and McKeachie (1979) contend that the use of student ratings increases the likelihood that excellence in teaching will be recognized and rewarded, which surely provides motivation for instructors and highlights another benefit of student evaluations. As for the criticism that students do not share with faculty the meaning of good teaching and, therefore, cannot accurately evaluate instructors, Feldman (1988) argues that students and teachers actually do agree generally on the components of effective teaching.

Undoubtedly, a sound argument can be made for the continued implementation of student evaluations of teaching in the higher learning. Nonetheless, those who oppose the use of SET's – especially for salary, promotion, and tenure decisions – constitute a strong and vocal group who gladly take to task the aforementioned perceived merits. In the following pages, the author turns his attention to the concerns of the detractors of student evaluations.

Perceived Disadvantages of SET's

Those who oppose the academy's complete reliance on student evaluations to measure teaching effectiveness are quick to point out what they see as one of the most problematic aspects of the method: No universal definition of effective teaching exists

(Monroe & Borzi, 1989; Spencer, 1992). One could argue that student outcomes constitute the strongest measure of teaching effectiveness; however, out of that definition arise complicated questions concerning what exactly students should learn in specific courses and, even more difficult to answer, how to adequately assess student learning.

Several studies have attempted to discover the dimensions of effective teaching. For example, Swartz, White, and Stuck (1990) identified as the factors of good teaching (1) clear instructional presentation and (2) management of student behavior. A few years later, Lowman and Mathie (1993) suggested that the characteristics of effective teaching are (1) intellectual excitement and (2) interpersonal rapport, while Brown and Atkins (1993) cited three characteristics of effective teachers: (1) caring, (2) systematic, and (3) stimulating. Moreover, Patrick and Smart (1998) posited that effective teachers must demonstrate (1) respect for students, (2) organization and presentation skills, and (3) the ability to challenge students. Still other scholars have cited as many as nine factors of effective teaching (Marsh & Dunkin, 1992).

Surely, detractors argue, an assessment form consisting of a few items that students rate on a five-point scale at the end of a semester cannot accurately measure the complexity and multidimensionality of effective teaching – especially considering educators, nor anyone else, for that matter, can agree on the components of effective teaching. Indeed, items included on typical student evaluation forms (such as “Rate the overall quality of this instructor based on the following variables: quality of speaking, clarity of objectives, and enthusiasm”) equate effective teaching with good in-class teaching behaviors (Wagenaar, 1995). And that is not good enough, say opponents of student evaluations. As Wagenaar (1995) states:

> Effective teaching is more than clear outlines written on the board and good speaking mannerisms. … Effective teaching also includes teaching students how to question assumptions, how to connect the course content with other content in the major and outside the major, how to learn to discover knowledge for themselves, how to create new wholes from discrete parts, how to use what is taught in their own lives as students and future citizens, how to work with other collaboratively, how to think in the manner of discipline, how to critique established ways of knowing, and the like. [Measuring teaching effectiveness] should move from examining only teacher behaviors to examining what and how well something is taught (p. 65).

Another concern surrounding the use of student evaluations of teaching centers on the fact that most rating forms are submitted anonymously. Critics question the practice of deciding issues of promotion, salary, and tenure based, at least in part, on anonymous student evaluations (Fries & McNinch, 2003). As a result, some have suggested requiring students to sign their evaluations with the assurance that their teachers will not have the opportunity to match up names with comments and scores (Baslow, 1995; Neath, 1996). The obvious rationale of this proposal is to promote personal accountability among students in the teaching evaluation process.

A number of scholars have studied the effects of signed and unsigned SET's and found that students tend to give higher ratings when they identify themselves compared to when they remain anonymous (Feldman, 1979; Blunt, 1991; and Fries & McNinch, 2003). Of course, a number of factors may influence this result, as Feldman (1979) concedes.

For instance, whether a student has received his or her grade prior to completing the evaluation, or whether a student believes a possibility of confrontation exists with the teacher in question certainly may bias the outcome of signed SET's. Additionally, the level to which students trust faculty and administrators to uphold promises of confidentiality directly influence the integrity of signed evaluations. Considering all of the factors that may prompt students to give unrealistically high scores to their instructors, most scholars agree that SET's should remain anonymous despite the risks inherent in anonymity (Centra, 1993; McCallum, 1984).

Unquestionably, the most common criticism of student evaluations involves the many types of biases that, opponents say, skew the results. In the following pages, the author explores three categories of bias that critics argue must be remedied before student evaluations of teaching can be deemed valid and reliable: instructor characteristics, student characteristics, and course characteristics.

Instructor Characteristics

In a study of the Web site RatemyProfessors.com, Felton, Mitchell, and Stinson (2004) found that students gave the highest ratings not to instructors who were the most helpful or clear in their teaching, or even from whom they learned the most. Instead, students using the Web-based evaluation gave the highest marks to instructors they deemed "hot" or good looking. Furthermore, the study concluded that easiness was the second most-likely factor to merit a teacher a positive rating (Felton, Mitchell, & Stinson, 2004). Not surprisingly, many professors (especially, one would assume, those who don't fit into either of the aforementioned categories of "hot" and "easy") and other detractors

criticize the ridiculousness of bias that, they argue, is prevalent in student evaluations of teaching.

Remarkably, Felton, Mitchell, and Stinson's (2004) study is not alone in its conclusions concerning instructor characteristics that influence student evaluation results. Weinberg, Fleisher, and Hashimoto (2007) conducted one of the largest and most ambitious studies on SET's. (The researchers analyzed about 50,000 student evaluations in 400 economics courses over a period of several years.) The authors concluded that the grades students receive (or expect to receive) in a course correlate with the scores students give the instructors. In other words, students are rewarding those teachers who reward them with good grades (Weinberg, Fleisher, and Hashimoto, 2007). This determination is supported by numerous studies (Braskamp & Ory, 1994; Centra, 1979; Marsh & Dunkin, 1992). Understandably, some academics worry that this trend is leading to grade inflation among instructors who wish to receive high student evaluation scores.

Although Marsh and Dunkin (1992) hypothesize that the grades/ratings correlation could be a result of instructors' setting more lenient grading standards in an effort to receive better evaluations, the scholars offer another plausible explanation: They hypothesize that more effective instructors cause students to work harder, learn more, and earn better grades. Thus, Marsh and Dunkin (1992) contend the relationship between expected grades and teacher ratings supports the validity of SET's. While their alternative explanation is optimistic, many researchers hold to the more realistic theory: Instructors who are easy – and not necessarily the best teachers – are receiving positive ratings

(Chacko, 1983; Koshland, 1991; Nimmer & Stone, 1991; Weinberg, Fleisher, & Hashimoto, 2007).

In addition to physical appearance and leniency, critics argue that gender and race also contribute to biases that negate the validity of SET's. Despite this criticism, research that investigates gender bias in student evaluations of teaching has produced mixed results. Some studies have found few differences between evaluations of male and female instructors on the basis of gender alone (Basow & Howe, 1987; Feldman, 1992; Harris, 1975). However, others have found more significant gender bias in that male students rate female teachers lower than male teachers (Etaugh & Riley, 1983; Kaschak, 1978). Again, little consensus has been reached concerning gender bias. However, Weinberg, Fleisher, and Hashimoto (2007) reported that students – controlling for other factors – tended to give lower scores to foreign-born instructors despite not finding any correlation between instructor identity and the level of learning that took place.

Another perceived bias that SET opponents criticize, but that research hardly has verified, involves the "Dr. Fox" effect. The concern focuses on an instructor's entertainment level, which Naftulin, Ware, & Donnelly (1973) concluded influences student evaluation scores. In their famous study, Naftulin et al. placed an actor, known as "Dr. Fox," in a college classroom where he presented a highly entertaining lecture that included no substantive content. The actor received rave student evaluation scores, which led the researchers to determine that highly charismatic lecturers can seduce students into giving high ratings despite learning nothing.

More than 30 years after the original study, the "Dr. Fox" effect has been harshly criticized and receives very few supporters in current literature (Perry, 1990). As Marsh

and Ware (1982) discovered, when students are not given incentive to learn – through grades, for example – as they were not in the "Dr. Fox" study, the entertainment level of a professor had a much greater affect on student ratings than the content presented. On the other hand, when students do have incentive to learn – as they typically do in a normal classroom setting – the entertainment level of an instructor was less important, and the "Dr. Fox" effect was essentially nonexistent.

Student Characteristics

Of all the student characteristics that have been studied in an effort to discover potential biases in SET's, a student's expectation of a course and its instructor is the single most important factor that influences teacher evaluations (McKeachie, 1979). Essentially, the research in this area has reaffirmed the self-fulfilling prophecy concept: Students who expect an instructor to be good typically finds this to be true. Furthermore, these students who hold high expectations generally rate their instructors higher than those with lower expectations (Koermer & Petelle, 1991). The obvious criticism among opponents of SET's is that a student's expectation level is outside the control of the instructor, and, therefore, the bias skews teacher assessment results.

Researchers also have concluded that the emotional state a student is in when he or she completes a teacher rating form affects the validity of the results. Small, Hollenbeck, and Haley (1982) found that the more anxious, depressed, frustrated, and hostile students were at the end of a semester, the more likely they were to give poor scores when evaluating teachers. Similar to the perceived bias of student expectation, critics of SET's argue that instructors cannot control their students' emotional states, and,

as a result, the potential for unfair bias detracts from the validity of traditional teacher evaluations.

Course Characteristics

Other background variables that critics believe bias student evaluations of teaching include class time, class size, subject area, and course workload. For example, Koushki and Kuhn (1982) concluded that instructors who taught classes that met early in the morning, shortly after lunch, or in the late afternoon received generally lower student evaluation scores. Moreover, in the same study, the researchers found class time influenced SET scores more than any other variable, including student gender, year in school, field of study, and expected grade.

More extensive research has addressed the variable of class size and how it affects the validity of student evaluations of teaching. Unquestionably, the literature supports the claim that smaller classes tend to give higher teacher evaluation ratings (Feldman, 1978; McKeachie, 1990). Interestingly, Scott (1977) determined that teachers who believed their classes were too large for them to present the material effectively received lower ratings than other instructors. This finding leads some to hypothesize that how instructors feel about the size of their classes may affect their performance and, accordingly, their evaluation scores (Feldman, 1978).

Like class size, scholars have concluded that the variable of subject matter also influences SET ratings. Research shows that mathematics and the sciences boast the lowest student evaluation scores (Cashin, 1992; Centra & Creech, 1976). In fact, the discrepancy is so great that some, like Ramsden (1991), have argued that student ratings should not be compared across disciplines. Centra (1993) posits that instructors in

mathematics and the sciences receive low scores because (1) they are less student-oriented, (2) they are required to spend more time in research, and (3) their courses are faster paced. Along these lines, Centra (1993) contends that students and teachers in the natural sciences have vastly different ideas about what constitutes appropriate pace and workload in a course, which certainly can influence SET results.

Conclusion: Suggestions for Improvement

It has been the purpose of this paper to survey the landscape and discuss the perceived merits and disadvantages of student evaluations of teaching effectiveness. Whether one supports or opposes the use SET's for salary, promotion, and tenure decisions, surely everyone can agree that the assessment method can be improved. As mentioned earlier, the academy has not agreed upon a single definition of effective teaching. Even if it did, the definition may not correspond with students' definitions of quality instruction. As a result, involving students in the creation of teacher rating forms may be a helpful step. Additional research also should investigate whether students are able to accurately measure a teacher's effectiveness, and not simply his or her likability level, or the likability level of the course itself. Furthermore, extensive analysis of how instructors use SET feedback to improve their teaching would be worthwhile.

Scholars have noted that the best way to evaluate a teacher's effectiveness is to measure how much students have learned in his or her class. Put plainly, teachers whose students learn a lot are good instructors; teachers whose students learn little are ineffective instructors. Of course, as discussed in this paper, adequately measuring what a student learns is no easy task. Pre- and post-tests based on instructors' course objectives seemingly would work well, but, as Murray (2005) concedes, this type of assessment is

plausible only in academic disciplines where a standardized test assesses basic student knowledge of subject matter.

Although research has not concluded whether requiring students to sign their evaluations actually leads them to more thoughtful consideration of teaching effectiveness, the idea has some merit. If nothing else, signed rating forms would allow administrators to randomly select students with whom to conduct interviews; in-depth interviews with students surely would provide insight into an instructor's strengths and weaknesses that quantitative rating forms cannot provide. As previously noted, signed SET's also would add a level of personal responsibility on the part of the student that does not exist with unsigned rating forms.

Perhaps the most important aspect of appropriately measuring teaching effectiveness is triangulation. In other words, the multidimensionality of teaching must be assessed through multidimensional methods – specifically, peer review. As Weinberg, Fleisher, & Hashimoto (2007) assert, student evaluations of any form are best used in conjunction with peer reviews of teaching. Moreover, Palmer (1998) highlights the need for peer assessment:

> Though we teach in front of students, we almost always teach solo, out of collegial sight… We pay a high price for this privatization. Consider the way teaching is evaluated. When we cannot observe each other's teaching, we get evaluation practices that are distanced, demoralizing, and even disreputable. Lacking firsthand information about each other's work, we allow the artifacts of the student survey to replace the facts that can be known only in person (p. 142).

Peer evaluation could come in the form of colleagues' observing a teacher's performance during a class session, or through portfolio reviews. Requiring teachers to compile throughout a semester a portfolio that represents all aspects of their teaching would provide administrators or outside reviewers a solid basis for assessment.

Finally, self-evaluation is crucial for instructors wanting to improve their classroom teaching. However, as Wagenaar (1995) contends, academic departments must assist faculty in this endeavor. Administrators must take seriously the importance of faculty goal-setting, and they should provide the resources necessary to enable instructors to achieve one of the most significant goals in the higher learning – effective teaching.

References

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

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.

Avi-Itzhak, T., & Kremer, L. (1986). An investigation into the relationship between university faculty attitudes toward student rating and organizational and background factors. *Educational Research Quarterly, 10*, 31-38.

Basow, S. A. (1995). Student evaluations of college professors. *Journal of Educational Psychology, 87*(4), 656-665

Basow, S. A., & Howe, K. G. (1987). Evaluations of college professors: Effects of professors' sex-type, and sex, and students' sex. *Psychological Reports, 60*, 671-678.

Blunt, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. *Community College Review, 18,* 48-54.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work*. San Francisco: Jossey Bass.

Brown, G., & Atkins, M. (1993). *Effective teaching in higher education*. London: Routledge.

Cashin, W. E. (1988). Student ratings of teaching: A summary of the research. Manhattan, KS: Center for faculty Evaluation and Development, Kansas State University.

Cashin, W. E. (1992). Student ratings: The need for comparative data. *Instructional Evaluation and Faculty Development, 12*, 1-6.

Centra, J. A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal, 14*, 17-24.

Centra, J. A. (1979). *Determining faculty effectiveness*. San Francisco: Jossey-Bass.

Centra, J. A. (1993). *Reflective faculty evaluation*. San Francisco: Jossey-Bass.

Centra, J. A., & Creech, F. R. (1976). The relationship between student teachers and course characteristics and student ratings of teacher effectiveness. Project report 76-1, Princeton, NJ: Educational Testing Service.

Chacko, T. I. (1983). Student ratings of instruction: A function of grading standards. *Educational Research Quarterly, 8*, 19-25.

Cohen, P. A. (1980). Using student ratings feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*, 321-341.

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

Epstein, D. (2006, January 20). For all to see. *Inside Higher Ed*. Retrieved February 27, 2007, from http://insidehighered.com/news/2006/01/20/evals

Etaugh, C., & Riley, S. (1983). Evaluating competence of women and men: Effects of marital and parental status and occupational sex-typing. *Sex Roles, 9*, 943-952.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education, 9*, 199-242.

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education, 10*, 149-172.

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

Feldman, K. A. (1992). College students' views of male and female teachers: Part I-Evidence from the social laboratory and experiments. *Research in Higher Education, 33*, 317-351.

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness, and sexiness. *Assessment and Evaluation in Higher Education, 29*(1), 91-108.

Fries, C. J., & McNinch, R. J. (2003). Signed versus unsigned student evaluations of teaching: A comparison. *Teaching Sociology, 31*(3), 333-344.

Gallagher, T. J. (2000). Embracing student evaluations of teaching: A case study. *Teaching Sociology, 28*(2), 140-147.

Harris, M. B. (1975). Sex role stereotypes and teacher evaluations. *Journal of Educational Psychology, 67*, 751-756.

Kaschak, E. (1981). Another look at sex bias in students' evaluations of professors: Do winners get the recognition that they have been given? *Psychology of Women Quarterly, 5*, 767-772.

Koermer, C. D., Petelee, J. L. (1991). Expectancy violation and student rating of instruction. *Communication Quarterly, 39*, 341-350.

Koon, J., & Murray, H. G. (1995). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *Journal of Higher Education, 66*, 61-81.

Koshland, D. E. (1991). Teaching and research. *Science, 25*(1), 249.

Koushki, P. A., & Kuhn, H. A. J. (1982). How reliable are student evaluations of teachers? *Engineering Education, 72*, 362-367.

Lowman, J., & Mathie, V. A. (1993). What should graduate teaching assistants know about teaching? *Teaching of Psychology, 20*(2), 84-88.

Lewis, J. M., & Benson, D. E. (1998). Course evaluations. In J. Lewis (Eds.). *Tips for teaching introductory sociology*. Belmont, CA: Wadsworth.

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

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J.C. Smart (Ed.). *Higher education: Handbook of theory and research*. New York: Agathon Press.

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*(11), 1187-1197.

Marsh, H. W., & Ware, J. E. (1982). Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the 'Dr. Fox' effect. *Journal of Educational Psychology, 74*, 126-134.

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education, 21*, 150-158.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe, 65*, 384-397.

McKeachie, W. J. (1990). Research on college teaching: The historical background.

Journal of Educational Psychology, 82, 189-200.

Menges, R. J. (1991). The real world of teaching improvement: A faculty perspective. In M. Theall & J. Franklin (Eds.). *Effective practices for improving teaching: New directions for teaching and learning*. San Francisco: Jossey-Bass.

Monroe, C., & Borzi, M. G. (1989). Methodological issues regarding student evaluation of teachers: A pilot study. *ACA Bulletin, 70*, 73-89.

Murray, H. G. (2005). Student evaluation of teaching: Has it made a difference? In the Annual Meeting of the Society for Teaching and Learning in Higher Education, June 2005 (pp. 1-15). Charlottetown, Prince Edward Island, Canada.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48*, 630-635.

Neath, I. (1996). How to improve your teaching evaluations without improving your teaching. *Psychological Reports, 78*, 1363-1372.

Nimmer, J.G., & Stone, E. F. (1991). Effects of grading practices and time of rating on student ratings of faculty performance and student learning. *Research in Higher Education, 32*, 195-215.

Obenchain, K. M., Abernathy, T. V., & Wiest, L. R. (2001). The reliability of students' ratings of faculty teaching effectiveness. *College Teaching 49*(3), 100-104.

Overall, J. U., & Marsh, H. W. (1979). Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcome. *Journal of Educational Psychology, 72*, 321-325.

Palmer, P. J. (1998). *The courage to teach: Exploring the inner landscape of a teacher's*

life. San Francisco: Jossey-Bass.

Patrick, J., & Smart, R. M. (1998). An empirical evaluation of teacher effectiveness: The emergence of three critical factors. *Assessment and Evaluation in Higher Education, 23*(2), 165-178.

Perry, R. P. (1990). Introduction to the special section: Instruction in higher education. *Journal of Educational Psychology, 82*, 183-188.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The course experience questionnaire. *Studies in Higher Education, 16*, 129-150.

Read, W. J., Rama, D. V., Raghunandan, K. (2001). The relationship between student evaluations of teaching and faculty evaluations. *Journal of Education for Business, 76*(4), 189-192.

Scott, C. S. (1977). Student ratings and instructor-defined extenuating circumstances. *Journal of Educational Psychology, 69*, 744-747.

Seldin, P. (1993). The use and abuse of student ratings of professors. *Chronicle of Higher Education, 39*(46), p. A40.

Small, A. C., Hollenbeck, A. R., Haley, R. L. (1982). The effect of emotional state on student ratings of instructors. *Teaching of Psychology, 9*, 205-208.

Spencer, P. A. (1992). *Improving teacher evaluation*. Riverside, CA: Riverside Community College.

Swartz, C. W., White, K. P., Stuck, G. B. (1990). The factorial structure of the North Carolina Teacher Performance Appraisal Instrument. *Educational and Psychological Measurement, 50*(1), 175-185.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief

review. *Assessment and Evaluation in Higher Education, 23*(2), 191-212.

Wagenaar, T. C. (1995). Student evaluation of teaching: Some cautions and suggestions. *Teaching Sociology, 23*(1), 64-68.

Wallace, J. J., & Wallace, W. A. (1998). Why the costs of student evaluations have long since exceeded their value. *Issues in Accounting Education, May*, 443-448.

Weinberg, B. A., Fleisher, B. M., & Hashimoto, M. (2007). Evaluating methods for evaluating instruction: The case of higher education. National Bureau of Economic Research working paper 12844: Cambridge, MA.

Wright, R.E. (2006). Student evaluations of faculty: Concerns raised in the literature, and possible solutions. *College Student Journal, 40*(2), 417-422.