Assessment & Evaluation in Higher Education
Vol. 32, No. 4, August 2007, pp. 403–415

Toward a more comprehensive approach to evaluating teaching effectiveness: supplementing student evaluations of teaching with pre–post learning measures

Kimberly Stark-Wroblewski*, Robert F. Ahlering and Flannery M. Brill
Central Missouri State University, USA

The use of student evaluations of teaching (SETs) to assess teaching effectiveness remains controversial. Without clear guidelines regarding how to best document effective teaching, faculty members may wonder how to convincingly demonstrate teaching effectiveness in preparation for promotion and tenure review. Based on a study that examined the relations among student grades, learning, and SETs, we identify a relatively unencumbered approach to documenting teaching effectiveness more comprehensively than through the use of SETs alone. Students enrolled in eight sections of general psychology ($N = 165$) completed pre- and post- measures of learning, SETs, and a brief demographic questionnaire. Results of a regression analysis provided partial support for the notion that SETs and learning measures assess distinct aspects of teaching effectiveness. In preparing documentation for promotion and tenure review, faculty members should consider including measures of student learning along with SETs in order to document teaching effectiveness more convincingly and comprehensively.

Introduction

Student evaluations of teaching (SETs) as indicators of teaching effectiveness, although criticized and controversial, appear to be heavily used in important personnel decisions such as hiring, promotion and tenure. The use of SETs to assess teaching effectiveness is further complicated by the number of audiences involved (students, faculty and administrators) and the multidimensional nature of teaching.

*Corresponding author. Department of Psychology, Central Missouri State University, Warrensburg, MO, 64093, USA. Email: stark@cmsu1.cmsu.edu.

ISSN 0260-2938 (print)/ISSN 1469-297X (online)/07/040403–13

DOI: 10.1080/02602930600898536

These issues leave faculty in a quandary as to how to convincingly demonstrate effective teaching in preparation for promotion and tenure review. This paper will describe how pre–post measures of learning tailored to individual courses, when used to supplement data from SETs, may provide a partial solution. Results of a study examining the relations among student grades, learning and SETs provide further support for the notion that SETs and measures of learning assess distinct aspects of teaching effectiveness.

Criticisms of the use of SETs

Although SETs are widely used for both formative (i.e., to promote improvement in teaching) and summative (i.e., as a form of quality assurance) feedback, their use is not without controversy (see Greenwald, 1997; Marsh & Roche, 1997; Kulik, 2001; Olivares, 2003). In particular, concerns have been raised regarding the possible biasing effect of grade inflation (see Eiszler, 2002). In essence, it has been argued that faculty may be tempted to assign higher grades to students in exchange for higher (i.e., better) student evaluation ratings. Thus, grading leniency is seen as a contaminant or biasing variable that spuriously inflates student ratings of teachers who assign higher grades. Olivares (2003) has even argued that, with the widespread use of SETs, 'academic control and authority have been put in the hands of students' (pp. 242–243), thereby resulting in an increase in consumer-oriented teaching and a decrease in student learning.

Anthony Greenwald (1997), action editor on a current issues section of the *American Psychologist* devoted to the debate over validity concerns associated with the use of SETs, suggested that the commonly observed positive correlation between grades and student ratings, which usually ranges from .10 to .30 (Feldman, 1997), is a matter of discriminant (in)validity. After reviewing extensive research providing evidence for the convergent validity of SETs, he concluded that grading leniency remains a contaminant in the use of SETs as a way to assess teaching effectiveness. Likewise, after presenting five theories that might explain the grades-rating correlation, Greenwald and Gillmore (1997) concluded that the 'leniency' or 'grade satisfaction' theory, in which students assign high ratings in appreciation for lenient grading, best explains the patterns of relationships observed in prior research.

Multiple audiences for SETs and uncertainty regarding how SETs are utilized

As Spencer and Schmelkin (2002) have aptly pointed out, when it comes to the use of SETs, there are essentially three interested parties, faculty, students and administrators. Based on their review of the literature pertaining to students' perceptions of the use of SETs, they concluded that students are 'not too optimistic about the overall weight put by administrators and faculty on student opinion' (p. 398). Results of their own study of student perspectives regarding SETs suggested that, although students believe SETs *should* be used in administrative decision-making (e.g., promotion and tenure decisions), students are generally pessimistic that this actually occurs.

Spencer and Schmelkin concluded that although students wish to express their opinions and thereby influence teaching, many appear skeptical about the actual use of SETs. In another study of students' perceptions of the use of SETs, Worthington (2002) found that students who believe SETs will be used for *summative purposes* (i.e., in making decisions regarding promotion, tenure and salary) tend to assign *lower* ratings of teaching; in contrast, those who believe SETs will be used for formative purposes (i.e., in order to improve teaching in the future) tend to assign higher (i.e., more favourable) ratings. Depending on the extent to which SETs are used in making personnel decisions, these findings might be rather unsettling for the untenured faculty member. As noted by Spencer and Schmelkin, research is needed on how those in a position to use SETs for promotion and tenure decisions (i.e., administrators) actually use such data. In the absence of this information, faculty may feel at a loss when deciding how to best present evidence of teaching effectiveness for promotion and tenure review.

In our review of the literature, we were unable to locate any conclusive information regarding how, and to what extent, data from SETs are used in administrative decisions. In fact, the most declarative statement we were able to find on this issue came from the US Department of Education's National Center for Education Statistics report on Institutional Policies and Practices (Berger *et al.*, 2001), which noted, 'Measures based on student inputs or results were used by most institutions, with 86 per cent using at least one student-based measure to evaluate full-time faculty' (p. vi). Given the dearth of information available regarding how SETs are used for personnel decisions and the continued debate over the validity of SETs as a measure of teaching effectiveness, individual faculty members are left to wonder how heavily SET data will factor into important decisions affecting their careers (i.e., promotion and tenure), further contributing to the apprehension faculty members experience in preparing for the review process.

Multidimensional nature of teaching and recommendations for improving upon SETs

Olivares (2003) has argued that in order to assess the validity of SETs one must first arrive at an adequate definition of teaching effectiveness. However, a universally agreed-upon definition of effective teaching remains elusive. Marsh and colleagues (see Marsh, 1991; Marsh & Dunkin, 1992; Marsh & Roche, 1997) and others (see d'Appolonia & Abrami, 1997; Feldman, 1997) have noted that teaching is multidimensional in nature, and hence there are many possible indicators of effective teaching. For instance, in addition to examining student achievement (i.e., learning), other factors such as student motivation, interest in the subject matter, and career aspirations can be impacted by teaching. Marsh has argued that we need to consider whether specific dimensions of SETs are highly correlated with specific criteria 'to which they are most logically related' (p. 419). For instance, group interaction and individual rapport, two factors on the Students' Evaluation of Educational Quality (SEEQ) instrument (Marsh, 1983, 1984, 1987; Marsh & Dunkin, 1992) were found to be logically related to class size (see Marsh & Roche, 1997). Further studies investigating specific criteria assessed by SETs may prove useful in the debate over the validity of SETs.

While debate continues over the validity of SETs, and, more fundamentally, the definition of teaching effectiveness, what are individual faculty members, pressed to demonstrate their teaching effectiveness, to do in the meantime? Rindermann and Schofield (2001) have argued that objective criteria, such as test results, should be used to supplement student ratings of instruction. However, as noted by Greenwald and Gillmore (1997), in many instances alternatives to student ratings in assessing teaching effectiveness are not readily available or may be expensive and/or cumbersome to employ. Nonetheless, evaluating student learning outcomes is considered by many as an important part of evaluating teaching effectiveness (see Diamond, 1995; Kulik, 2001; Rindermann & Schofield, 2001; Olivares, 2003). Supplementing SET data with learning outcome data may also be in the best interest of the faculty member, particularly in light of Worthington's (2002) finding that students tend to assign lower SET ratings when they believe the data will be used for summative, rather than formative, purposes. Further, as Biggs (2003) has described in his constructive alignment model, good teaching aligns teaching methods with assessment in order to support student learning.

Pre–post measures of learning as a partial solution

Recently, Arthur *et al.* (2003) employed a pre–post assessment of student learning to examine the relations among student grades, learning (as measured by the pre–post assessment) and student evaluations of teaching. Their results revealed a medium relationship between student grades and learning as well as a small relationship between student ratings of teaching effectiveness and learning. The authors concluded that because student ratings of teaching effectiveness and learning measures appear to assess independent aspects of teaching effectiveness, both criteria may legitimately be used in conjunction with one another to obtain a more comprehensive assessment of teaching effectiveness. The pre–post assessment technique employed by Arthur *et al.* may present a viable method of objectively assessing student learning for some instructors. However, given the numerous demands on class time already in existence, many instructors may be reluctant to devote class time solely to the collection of such assessment data.

The genesis of the present study stemmed, in large part, from the first author's need to document teaching effectiveness, in the form of student learning, in preparation for promotion and tenure review. Specifically, the first author sought to document student learning in her introductory psychology (i.e., general psychology) course; however, the content addressed in this course does not correspond directly with learning outcomes assessed via available nationally standardized exams such as the major field tests (Educational Testing Service, 2004). Therefore, the first two authors, who both routinely teach general psychology, collaborated with one another in developing a method to assess student learning in this course. Using both empirical and rational criteria, we developed a 30-item pre-test of student learning by selecting six items from each of the five exams administered in the course. On the first day of class, students are asked to complete the pre-test, and these items are also used to (a) introduce students to the course by illustrating topic areas that will be covered during the semester; and (b) illustrate the types of questions that will be included on exams administered in the course. Post-test scores are calculated by selecting scores for the relevant items across each of the five exams and then summing these scores at the end of the semester. In this way, the post-test is essentially 'embedded into' exams that students complete in class, allowing us to examine student learning without usurping valuable class time. At the end of the semester pre- and post-test scores are compared to assess student learning in the course.

Aims and hypotheses of the current study

The present study had two aims. First, we sought to find a way to supplement the use of SETs as a measure of teaching effectiveness with a form of 'objective' student learning in a manner that would not expend considerable class time and in a way that might benefit our students. Second, we sought to contribute to the literature available regarding the validity of SETs by examining the relations among student learning, grades, and SETs. With regard to the second aim, we hypothesized that, in conjunction with the existing literature, there would be a small to moderate positive correlation between SET scores and grades. Second, similar to Arthur and colleagues (2003), we expected to find a small to medium positive correlation between learning (as measured by the difference in pre- and post-test scores) and grades. Third, and again in parallel with the findings of Arthur *et al.*, we hypothesized that there would be a small to medium positive correlation between learning and SET scores. Fourth, in consideration of Marsh's (1991) recommendation to investigate whether specific dimensions of SETs are logically related to certain criteria, we hypothesized that the relationship between learning and an item on the SET specifically inquiring about learning would be moderate to large. These four hypotheses were tested by performing Pearson Product Moment correlations among the major variables of interest—i.e., learning, SET scores and grades. Lastly, similar to Arthur *et al.*, we expected to find evidence that SETs and measures of learning assess distinct aspects of teaching effectiveness. This hypothesis was tested by performing a simultaneous regression analysis with grades as the criterion variable, and with learning and SET scores serving as predictor variables. We anticipated that learning scores would predict variance in grades above and beyond that explained by SET scores alone and vice versa.

Method

Students enrolled in eight sections of an introductory psychology (general psychology) course taught by two different instructors at a mid-sized university located in the Midwestern region of the US were invited to participate in the study. The number of students per section ranged from 41 to 62 ($M = 51.75$, $SD = 7.21$). On the first day of class all students were asked to complete a 30-item multiple-choice test ('pre-test') that addressed material from the entire course. Students were informed that the pre-test would not count toward their grade in the course, and that they should not expect to know the answers at the time of administration, as this pre-test was diagnostic of their level of knowledge prior to having taken the course. They were also told that the pre-test might be useful in providing them with (a) a sampling of the kinds of topics to be covered in the course; and (b) an illustration of the types of questions that would be included on subsequent exams in the course. Items for the test used in each course were selected by the instructor of that course based on both rational and empirical criteria. Using item analysis data from previous exams, items were selected on the basis of their ability to distinguish between low versus high scoring examinees; content assessed by the items was also considered in item selection. Specifically, items with a difficulty level between .425 and .825 were targeted for inclusion in the pre–post test, and redundant items (i.e., items assessing overlapping content) were not included in the pre–post test.

Subsequent to the administration of the pre-test, five exams were administered in the course, with six of the pre-test items embedded into each exam. The summary of these six items across all five exams constituted the 30-item post-test. Thus, administration of the post-test did not require any additional class time. Students completed the SET and a brief demographic questionnaire following the last class period prior to the final exam. No extra credit was offered for participation.

The SET measure consisted of 16 items. All but one of the items were from the measure used by Arthur *et al.* (2003) in their study of the relationship between reaction criteria (SETs) and learning criteria (objective test scores). A sixteenth item was added which read 'I learned in this course'. Participants were asked to respond on a 5-point Likert scale ranging from '1' ('strongly agree') to 5 ('strongly disagree') in evaluating the course and instructor. The instrument contained items inquiring about specific aspects of teaching, such as the instructor's ability to answer questions, and present material effectively as well as items relating to rapport, grading, enthusiasm, workload, availability and ability to generate student interest in the subject matter. In addition, the instrument contained items in which the student was asked to provide a more global rating (e.g., 'On the whole this was a good course'). We were particularly interested in the total score on the SET (calculated by summing individual responses across all 16 items) as well as item 15, the 'learning' item ('I learned in this course'). The alpha coefficient for the instrument, indicating internal consistency, was .95.

Results

Of the 414 students enrolled in the eight sections of general psychology, a total of 188 students elected to participate in the study, yielding a 45% participation rate. However, only data from the 165 students who completed the pre-test and all of the post-tests were included in the final analyses. Of these 165 participants, 101 (61.2%) were female and 64 (38.8%) were male. The average grade point in the course for those participating in the study was 2.96 (*SD* = .92), based on the standard 4-point grading scale (i.e., A = 4, B = 3, C = 2, D = 1 and F = 0).

A learning score for each participant was calculated by subtracting each student's pre-test score from his or her post-test score, thereby yielding a difference score. The average learning score was 13.42 ($SD = 4.70$). There was significant improvement in correct responses from pre- ($M = 8.48$, $SD = 2.56$) to post-test ($M = 21.90$, $SD = 4.79$) [$t(164) = 36.69$, $p < .001$, $d = 2.86$]. It is noteworthy that, according to Cohen (1992), effect sizes (*d*s) exceeding 0.80 are considered large.

Pearson intercorrelations for the major variables of interest are shown in Table 1. We hypothesized that there would be a small to moderate positive correlation between SET scores and grades. Student grades and total SET scores were significantly correlated ($r = .18$, $p = .02$) and this correlation fell within the anticipated small to moderate range. Second, we hypothesized a small to medium positive correlation between learning scores and grades. Learning scores were significantly correlated with grades ($r = .66$, $p < .001$). Although hypothesis two predicted a small to moderate positive correlation between learning and grades, this correlation was, in fact, large. Third, we hypothesized that there would be a small to medium positive correlation between learning and SET scores. The correlation between learning and SET scores was small and only marginally significant ($r = .15$, $p = .06$). Fourth, we hypothesized that the relationship between learning and an item on the SET specifically inquiring about learning would be moderate to large. The SET item most highly correlated with learning ($r = .20$, $p = .009$) was the item that specifically inquired about learning (i.e., 'I learned in this course'). Although this correlation was significant, it was not in the moderate to large range predicted by Hypothesis 4.

For our fifth hypothesis, we expected to find evidence, similar to Arthur *et al.* (2003), that SETs and measures of learning assess distinct aspects of teaching effectiveness. Thus, Hypothesis 5 predicted that (a) learning scores would account for variance in grades over and above that explained by SET scores; and (b) SET scores would account for variance in grades over and above that explained by learning scores. A simultaneous regression analysis was performed with grades as the criterion variable, and with learning and SET scores serving as predictor variables. Positive beta weights were expected for both predictor variables. The regression equation was significant, $R^2 = .44$, adjusted $R^2 = .43$, $F\ (2, 162) = 62.82$, $p < .001$. Together,

Table 1. Variable intercorrelations

	1	2	3	4	5	6	7
Pre-test	–						
Post-test	.30***	–					
Learning score	–.24**	.86***	–				
Grades	.31***	.81***	.66***	–			
SET total score	.04	.17*	.15	.18*	–		
SET-learned	.08	.24**	.20**	.29***	.78***	–	

Notes: Learning score refers to the difference in scores from pre- to post-test on the student learning measure. SET-Learned refers to scores on SET item stating, 'I learned in this course'. $N = 165$. *$p < .05$, **$p < .01$, *** $p < .001$ (2-tailed).

learning and SET scores accounted for 44% of the variance. Positive beta weights were found for both learning ($\beta = .643$) and SET ($\beta = .084$) scores. Learning scores served as a significant predictor, $t(163) = 10.79$, $p < .001$, but SET scores did not, $t(163) = 1.41$, $p = .16$. Learning scores predicted 40% of the unique variance in grades, whereas SET scores predicted less than 1% of the unique variance. Thus, Hypothesis 5 was only partially supported.

Discussion

Not surprisingly, results of the present study coincide with those of previous investigations examining the relationship between SET scores and grades. That is, our results revealed a small to moderate relationship between SET scores and grades as has been reported in the previous literature. In fact, the correlation of .18 found in the present study falls approximately midway between the commonly observed ranges of .10 to .30 reported by Feldman (1997).

As compared to the findings of Arthur and colleagues (2003), our data revealed a considerably larger relationship between our measure of learning and grades. This discrepancy could be due to methodological differences between the two studies. In the present study, the post-test was embedded into existing exams administered throughout the semester, whereas Arthur et al. administered a post-test, independent of the exams, at the end of the course. Thus, in comparison to the pre–post test method used by Arthur *et al.*, our post-test occurred in closer temporal proximity to the actual coverage of course material. Second, in the present study, the post-test represented a subset of the exams, which, in turn, constituted a major portion of students' grades in the course. Students are undoubtedly more motivated to perform better on a post-test that contributes toward their final grades as compared to a post-test that is independent of course grades. While it could be argued that the magnitude of this relationship is spuriously inflated by the fact that the post-test scores constituted a portion of the student grades, it should be noted that, for all sections, the post-test scores in reality represented approximately 10% or less of students' grades. This seems unlikely to substantially influence the results of the present study. Moreover, this 10% estimate is probably an overstatement of the influence of the post-test scores on grades, since the statistical analyses included the difference score from pre- to post-test, rather than just the post-test score. Finally, in contrast to the Arthur *et al.* study, in the present study, post-test items were tailored to be course-specific and were chosen based on their ability to discriminate between high versus low performers on exams. These methodological differences might account for the stronger relationship found between learning and grades in our study as compared to the prior work of Arthur *et al.*

Our finding of a small and only marginally significant correlation between learning and SET scores (.15) might provide further evidence that, although these dimensions are somewhat related, they are largely independent of one another. Thus, it would appear that SETs and learning outcome measures assess distinct aspects of teaching effectiveness.

Although we found a small (.20), rather than the hypothesized moderate to large, correlation between learning and the specific SET item inquiring about learning, it is noteworthy that this particular item was the one most highly correlated with learning. This finding is in keeping with the suggestion of Marsh (1991) that researchers examine the extent to which specific dimensions of SETs are correlated with criteria with which they would most logically be expected to be related. Thus, it makes sense that the item inquiring specifically about learning ('I learned in this course'), rather than items assessing other aspects of teaching, such as rapport (e.g., 'The instructor seemed to care whether the students learned'), would correlate most highly with our objective measure of learning.

Similar to Arthur *et al.* (2003), we found at least partial support for the notion that SETs and learning measures assess distinct aspects of teaching effectiveness. That is, although both SET and learning scores were positively correlated with grades, only learning scores were predictive of the variance in grades over and above that explained by SET scores alone, whereas the reverse was not found to be true. While some might argue that these findings indicate there is little utility in using SETs to assess teaching effectiveness, we would argue for a more cautious interpretation and concur with Arthur and colleagues (2003) that both student learning and student ratings of teaching should be considered in evaluating teaching effectiveness. Arthur *et al.* have provided a useful conceptualization by drawing upon the organizational training evaluation and effectiveness literature, which distinguishes between reaction and learning criteria (Kirkpatrick, 1976). In applying this analogy to the debate over the use of SETs, Arthur and colleagues have noted, 'a distinction can be drawn between students' affective responses to the instructor or course (reaction criteria) and the amount of student learning that occurred (learning criteria)' (pp. 276–277). They have further suggested that assessment of teaching effectiveness depends on the goals of teaching, noting that if one 'considers teaching as a service and students as consumers of this service' then students' perceptions of the service rendered are 'the most relevant criteria' (p. 278). In contrast, one might evaluate teaching effectiveness based on how much students have learned, in which case, 'some pre/post measure of learning would be the most relevant criterion' (p. 278). We would argue that, at a minimum, *both* of these aspects of teaching need to be assessed in order to approach a more comprehensive method of evaluating teaching effectiveness.

As noted by d'Apollonia and Abrami (1997) teaching should be considered multidimensional in nature, and, as a result, more comprehensive systems for evaluating instruction need to be developed. Thus, although SETs may capture important aspects of teaching effectiveness, such as satisfaction with the course, instructors should consider supplementing SETs with other measures of teaching effectiveness. Because student learning has been cited by many authors as an important element of effective teaching (see Diamond, 1995; Kulik, 2001; Rindermann & Schofield, 2001; Olivares, 2003), instructors should consider including measures of student learning in their teaching evaluation repertoire. The embedded pre–post test method offers a way to assess student learning that is relatively efficient to administer (i.e., does not require much additional class time) and has the added benefits of providing students with a sample of exam questions and a quick overview of the course, while simultaneously providing instructors with a pre-course assessment of students' familiarity with the subject matter. The pre–post assessment method employed in the present study is unique in that the post-test was embedded into existing exams, thereby allowing the authors to generate course-specific evaluation measures and evaluate student learning without administering a separate post-test measure. Other instructors might also find it useful to develop individualized pre–post measures that best match the specific learning objectives in courses they teach.

One might argue that an individualized (i.e., instructor generated) pre–post test approach to assessing learning would complicate the review process, since such individualized assessments are, by their nature, not standardized and thus, data based on such approaches do not readily lend themselves for comparison. In other words, reviewers might wonder how to assess 'how much learning is adequate?' since no standard metric exists for such idiosyncratic measures. However, a simple solution would be for the individual instructor to report the *d*-value (i.e., effect size), related to the difference in scores from pre- to post-test as an indicator of the degree of student learning that occurred in a given course. In this manner, instructors can develop learning measures tailored specifically to individual courses, and yet still present a standard metric by which learning can be assessed and evaluated. Taken to an extreme, teaching effectiveness could be compared across faculty members (even across disciplines) by comparing instructors' respective *d*-values. This approach would appear to hold a great deal of promise for evaluating teaching effectiveness in a standardized fashion. We would caution against relying solely on such an approach, however, and again reiterate our stance that teaching, and hence its assessment and evaluation, is multidimensional in nature, thereby arguing for a more comprehensive approach to evaluating teaching effectiveness.

Limitations and suggestions for future research

It should be noted that findings based on this study might not generalize beyond the limits of our sample of college students enrolled in an introductory psychology course at a mid-sized university in the Midwestern US. Additionally, our sample may represent a selection bias in that students enrolled in select sections of introductory psychology self-selected to participate. Finally, the data reported here were generated from students enrolled in the courses of only two instructors. Thus, generalizations from the present findings may be limited in several ways.

It should also be noted that the present study did not actually measure 'teaching effectiveness' per se as this remains a theoretical, and presumably multidimensional, construct. Hence, no universally agreed-upon, comprehensive measure of teaching effectiveness currently exists. Rather, in the present study, student grades were used as a sort of 'proxy' measure of teaching effectiveness. In reality, of course, grades alone do not constitute an adequate measure of teaching effectiveness. Future studies should be directed at assessing other aspects of teaching effectiveness, such as those noted by Marsh and colleagues (see Marsh, 1991; Marsh & Dunkin 1992; Marsh &

Roche, 1997) and other authors (see d'Appolonia & Abrami, 1997; Feldman, 1997). For instance, future research could be directed at assessing the extent to which SET scores are predictive of other teaching-related outcomes (e.g., student motivation, students' decisions to major in particular disciplines, etc.).

Despite the limitations of the present study, these results contribute to the discussion over the use of SETs and lend support for utilizing a multifaceted approach to instructional evaluation. Given that the definition of effective teaching remains elusive, it is not surprising that the validity of SETs as a measure of teaching effectiveness continues to be debated (Olivares, 2003). Nonetheless, SETs continue to be widely used in evaluating teaching and, subsequently, in making important personnel decisions. The extent to which individual institutions and administrators rely on SET data in making personnel decisions remains relatively unknown. This state of affairs may leave faculty members wondering how to appropriately document teaching effectiveness. The pre–post method described herein presents a viable method of supplementing SET scores with learning outcome measures in an efficient manner that provides several instructional benefits. Additionally, individual instructors can present a standard metric, in the form of a *d*-score, as an indicator of student learning. As the debate over the use of SETs continues, and in the absence of clear guidelines for documenting teaching effectiveness, individual faculty members are encouraged to consider supplementing SETs with measures of student learning such as the pre-post assessment method described herein.[1]

Acknowledgement

We wish to thank our colleague, Dr. David S. Kreiner of Central Missouri State University, for his helpful comments on previous drafts of this manuscript.

Note

[1] An earlier poster based on this study was presented in 2004 at the 11th Annual Teaching Institute of the American Psychological Society, Chicago, Illinois, USA.

Notes on contributors

Kim Stark-Wroblewski is Associate Professor of Psychology at Central Missouri State University. Her Ph.D. is in counseling psychology. She frequently collaborates with students on research related to the scholarship of teaching.

Robert Ahlering is a faculty member within the Department of Psychology at Central Missouri State University. His Ph.D. is in the area of social/personality with some postdoctoral training in industrial/organizational psychology. His research has generally focused on attitudes.

Flannery Brill is a graduate student in the psychology program at Central Missouri State University. She is currently working on a masters thesis examining instructors' perceptions of online courses.

References

Arthur, W. Jr., Tubre, T., Paul, D. S. & Edens, P. S. (2003) Teaching effectiveness: the relationship between reaction and learning evaluation criteria, *Educational Psychology,* 23(3), 275–285.

Berger, A., Kirshstein, R., Rowe, E. & Zimbler, L. (2001) *Institutional policies and practices: results from the 1999 national study of postsecondary faculty, institution survey* (Washington, DC, US Department of Education). Available online at: www.nces.ed.gov/pubs2001/2001201.pdf (accessed 1 March 2006).

Biggs, J. B. (2003) *Teaching for quality learning at university: what the student does* (Buckingham, Society for Research into Higher Education).

Cohen, J. (1992) A power primer, *Psychological Bulletin,* 112(1), 155–159.

d'Apollonia, S. & Abrami, P. C. (1997) Navigating student ratings of instruction, *American Psychologist,* 52(11), 1198–1208.

Diamond, R. M. (1995) *Preparing for promotion and tenure review: a faculty guide* (Bolton, MA, Anker Publishing Company).

Educational Testing Service (2004) *Major field tests: test administration manual* (Princeton, NJ, Author).

Eiszler, C. F. (2002) College students' evaluations of teaching and grade inflation, *Research in Higher Education,* 43(4), 483–501.

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* (New York, NY, Agathon Press), 368–395.

Greenwald, A. G. (1997) Validity concerns and usefulness of student ratings of instruction, *American Psychologist,* 52(11), 1182–1186.

Greenwald, A. G. & Gillmore, G. M. (1997) Grading leniency is a removable contaminant of student ratings, *American Psychologist,* 52(11), 1209–1217.

Kirkpatrick, D.L. (1976) Evaluation of training, in: R. L. Craig (Ed.) *Training and development handbook: a guide to human resource development* (New York, NY, McGraw-Hill), 301–319.

Kulik, J. A. (2001) Student ratings: validity, utility, and controversy, in: M. Theall, P. C. Abrami & L.A. Mets (Eds) *The student ratings debate: are they valid? how can we best use them?* (San Francisco, CA, Jossey-Bass), 9–25.

Marsh, H. W. (1983) Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics, *Journal of Educational Psychology,* 75, 150–166.

Marsh, H. W. (1984) Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility, *Journal of Educational Psychology,* 76(5), 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(3), 253–388.

Marsh, H. W. (1991) A multidimensional perspective on students' evaluations of teaching effectiveness: reply to Abrami and d'Apollonia (1991), *Journal of Educational Psychology,* 83(3), 416–421.

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, NY, Agathon), 143–234.

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.

Olivares, O. J. (2003) A conceptual and analytic critique of student ratings of teachers in the USA with implications for teacher effectiveness and student learning, *Teaching in Higher Education,* 8(2), 233–245.

Rindermann, H. & Schofield, N. (2001) Generalizability of multidimensional student ratings of university instruction across courses and teachers, *Research in Higher Education,* 42(4), 377–399.

Spencer, K. J. & Schmelkin, L. P. (2002) Student perspectives on teaching and its evaluation, *Assessment and Evaluation in Higher Education,* 27(5), 397–409.

Worthington, A. C. (2002) The impact of student perceptions and characteristics on teaching evaluations: a case study in finance education, *Assessment and Evaluation in Higher Education,* 27(1), 49–64.