Return to academic standards: a critique of student evaluations of teaching effectiveness

Charles R. Emery
Tracy R. Kramer and
Robert G. Tian

The authors

Charles R. Emery is Assistant Professor of Management, Lander University, Greenwood, South Carolina, USA. **Tracy R. Kramer** and **Robert G. Tian** are Associate Professors of Business Administration, both at Erskine College, Due West, South Carolina, USA.

Keywords

Teachers, Evaluation, Performance, Effectiveness, Students

Abstract

A student evaluation of teaching effectiveness (SETE) is often the most influential information in promotion and tenure decision at colleges and universities focused on teaching. Unfortunately, this instrument often fails to capture the lecturer's ability to foster the creation of learning and to serve as a tool for improving instruction. In fact, it often serves as a disincentive to introducing rigour. This paper performs a qualitative (e.g. case studies) and quantitative (e.g. empirical research) literature review of student evaluations as a measure of teaching effectiveness. Problems are highlighted and suggestions offered to improve SETEs and to refocus teaching effectiveness on outcome-based academic standards.

Electronic access

The Emerald Research Register for this journal is available at
http://www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/0968-4883.htm

Quality Assurance in Education
Volume 11 · Number 1 · 2003 · pp. 37-46
 · ISSN 0968-4883
DOI 10.1108/09684880310462074

Background

A current practice among colleges and universities in the USA is for the administration to use a student evaluation instrument of teaching effectiveness as part of the faculty member's performance evaluation. In a study that tracked the use of student evaluations of faculty in 600 colleges between 1973 and 1993, Seldin (1993) found that the use of student evaluations of teaching effectiveness (SETE) increased from 29 percent to 86 percent. If these instruments are used in isolation, as they frequently are, and without alternative or collaborative measures, then students become the primary determinant of a lecturer's success or failure in his or her academic career. At institutions that emphasize teaching (as opposed to research), higher-than-average levels of teaching effectiveness are often expected. Therefore, it follows that student measurements of teaching effectiveness have the potential to buoy or sink a lecturer's career. When students are empowered to yield this much influence over the career's of their lecturers, combined with the demands on the lecturers for increasing course loads, student enrollments and student satisfaction, the long-term results may very well be an overall reduction in program quality.

One school of thought at many business schools is that students should be viewed as the products of the business program, rather than its customers (Emery *et al.*, 2001). In other words, the lecturers are the immediate customers and industry/society is the ultimate customer. From this position, it is clear that the use of SETE, which implicitly captures lecturer popularity, is inappropriate for measuring instructional effectiveness (i.e. learning). Ironically, while business departments purport to use student appraisals to increase total quality, Deming (1986) has suggested that the practice is inaccurate and demoralizing.

In addition to criticisms of the evaluation philosophy and the validity of the instrument, there is reason to criticize the use SETE as the only method of evaluating teaching effectiveness. Comm and Mathaisel (1998) observed that in some industries, subordinates are used to evaluate their bosses but *never* as the only measure of supervisor effectiveness. Typically, this is used as the least weighted of several methods to ascertain

Quality Assurance in Education
Volume 11 · Number 1 · 2003 · 37-46

the administrative ability of the manager. Conversely, the majority of business schools use it as either the only method of teaching effectiveness or the most heavily weighted method (Abrami *et al.*, 1990). In short, does the existing evaluation practice encourage business school faculty to teach their students with future employers in mind, or does it encourage faculty to teach with their own evaluations in mind? This paper posits that many of the current methods of evaluation do not meet the overarching educational objective of improved student learning. In order to examine this proposition, a review was conducted of both the qualitative (e.g. case analyses) and quantitative (e.g. empirical research) literature.

Literature review

From very beginning, student instructional rating questionnaires have been touted as a cheap and convenient means of evaluating the teaching of college and university faculty. College administrators eagerly embraced SETE in the 1960s because they were perceived to be able to offer a ready vehicle for assessing faculty hired to teach the droves of students entering post-secondary institutes. The perceived promise, technical appearance and utter simplicity of SETE have ensured the popular use of student instructional ratings for nearly 40 years now. Research, however, indicates that SETE is not the only possible source of information about teaching effectiveness, and it is certainly not the best source of that information. Nationally, researchers have conducted hundreds of academic exercises on the reliability and validity of SETE. We will summarize some of the findings from that research and add several anecdotal cases to make this academic review more personal.

Popularity and personality contests

It is widely believed that SETE is only a popularity contest that has little to do with learning. Dooris' (1997) research supports such a hypothesis in her review of student ratings and grades in large, multiple section courses (such as introductory chemistry or physics) taught by several instructors who use common textbooks and give identical examinations. Similarly, Arreola (1995), Aleamoni (1987), Feldman (1978), and Theal and Franklin (1990) reached the same conclusions after reviewing hundreds of studies dating back to the 1920s. Emery (1995) found in a study of 2,673 students at a major state university that instructors who bring food to class receive the highest ratings of teaching effectiveness. Abrami *et al.* (1982) suggest that instructional ratings should not be used in decision making about faculty promotion and tenure, because charismatic and enthusiastic faculty can receive favorable student ratings regardless of how well they know their subject matter. Further, these instructor attributes were not related to how much their students learned.

A meta-analysis of a dozen of these studies revealed "instructor expressiveness had a substantial impact on student ratings, but a small impact on student achievement" (Abrami *et al.*, 1982). Feldman (1986) found when the assessment is based on the perceptions of students or colleagues, the overall relationship of instructor personality to student ratings is substantial, with positive correlations ranging from moderate to high. Similarly, Jones (1989) examined the question of whether students can validly judge teaching effectiveness without having their ratings distorted by irrelevant contextual variables. The results of this study indicated that student ratings of a teacher's personality and teaching competence are significantly related, even when students have been alerted to the "irrelevance" of personality characteristics in evaluating teaching. While these studies suggest that students significantly link personal characteristics with teaching competence, research indicates that there is only a small positive correlation between a student's self-reported learning and affection for the teacher (Cashin, 1989).

Student achievement

Few would argue with the notion that measuring student achievement is the purest form of assessing teacher effectiveness. Most investigations, however, found little correlation between achievement and student ratings. For example, in a well-controlled meta-analysis, Cohen (1983) found that student achievement accounted for 14.4 percent of overall instructor rating variance. Other analyses have turned up even lower estimates of student rating validity. In a meta-analysis of 14 multi-section validity studies, McCallum (1984) found that student

Quality Assurance in Education
Volume 11 · Number 1 · 2003 · 37-46

achievement explained 10.1 percent and 6.4 percent (respectively) of overall instructor and course rating variance. And, in a quantitative analysis of six validity studies chosen for their exceptional control of student presage variables, Dowell and Neal (1982) found that student achievement accounted for only 3.9 percent of between-teacher student rating variance. Finally, in a more comprehensive study, Damron (1996) found that it is likely that most of the factors contributing to student instructional ratings are unrelated to an instructor's ability to promote student learning. This is particularly relevant considering that validity research indicated only marginal and unstable relationships between student ratings and instructional outcomes. Damron further suggests that since lecture content contributes much less to student instructional ratings, the price instructors pay for this strategy is lower student ratings and, possibly, loss of promotions, salary increments, or employment.

Situational factors and validity

Researchers have found that the validity of SETE can also be affected by situational factors or biases. For example, Dowell and Neal (1982) pointed out that the variability in validity coefficients, even in studies with reasonable methodological requirements, led them to suspect that the validity of student ratings is influenced by situational factors to an extent that a meaningful, generalizable estimate of their validity does not exist. "In general . . . no meaningful estimate of the validity of student ratings can be provided with confidence that is generalizable enough to be useful" (Dowell and Neal, 1982, p. 60). Abrami *et al.* (1990) draw a similar conclusion about variability in validity outcomes across validity studies and over rating dimensions. They indicate: "Whereas the average validity coefficient for global ratings is moderately positive, the results of these studies appear inconsistent both from study to study and across rating dimensions" (Abrami *et al.*, 1990, p. 230). Similarly, Dowell and Neal (1983) concluded that the situational variables so thoroughly contaminate the validity of self-reported student learning and teacher effectiveness indices that they can only be regarded as indices of "consumer satisfaction".

A common practice of administrators at small liberal arts colleges is to compare individual faculty ratings regardless of the teaching levels and disciplines. Research, however, indicates that cross-discipline ratings bias makes the validity of SETE questionable. For example, Cashin (1990) examined very large databases of students' ratings and found sizable differences in how students rate teaching across various academic disciplines. For instances, the high group tends to consist of the arts and humanities; English language and literature and history both fall into the medium-low group; the low groups tend to consist mostly of business, economics, computer science, math, physical sciences, and engineering; the biological and social sciences and health and other professions tend to fall somewhere in the middle. Aleamoni (1989) observed a similar pattern of variability regarding rating biases against required courses and student biases associated with various course levels, such as freshman, sophomore, and the like. He indicated that the variables that distinguish a required course from an elective, and that identify courses by level (freshman, sophomore, and so on) have generated significant differences in student ratings. For instance, the more students in a class taking the course as a requirement, the lower the overall rating will be. Moreover, freshmen tend to rate their teachers significantly lower than do sophomores; sophomores tend to rate them significantly lower than do juniors, and so on.

In short, there are three fundamental reasons to account the validity problems of SETE (Damron, 1996). First, validation studies that do not properly control for biasing factors (e.g. student characteristics, instructor characteristics, class characteristics) yield internally invalid and uninterpretable estimates of rating validity. Second, when appropriate controls are implemented, resulting validity estimates account for only a small fragment of between instructors rating variance. The proportion of variance accounted for appears to be inversely related to the scope of the controls. Third, even among well-designed validity studies, validity coefficients tend to be highly variable and mediated by situational factors to such a degree that coherent context-independent estimates of validity are not possible. The latter two problems have weighty implications

for the accuracy and developmental utility of student ratings.

User error

Damron (1996) suggests that even if a sufficiently valid rating questionnaire existed, there are no guarantees that interpretations of ratings data will be valid or consistent (or reasonable, coherent or fair). Further, Franklin and Theall (1990) observed that the problem of unskilled users, making decisions based on invalid interpretations of ambiguous or bad data, need to be carefully noticed. They note that ratings are particularly subject to sampling problems; the fact that classes with fewer than 30 students are statistically small samples means that special statistical methods are required for some purposes. According to Franklin and Theall, because of such problems, three types of mistakes are not uncommon in terms of SETE interpretation.

The first mistake is the interpretation of severely flawed data, with no recognition of the limitations imposed by problems in data collection, sampling, or analysis. In this case, misinterpretation of statistics could lead to a decision favoring one instructor over another, when in fact the two instructors or not significantly different. The second type of error occurs when, given adequate data, there is a failure to distinguish significant differences from insignificant differences. In this case, failure to use data from available reports may be prejudicial to an instructor whose performance has been outstanding but who, as a result of the error, is not appropriately rewarded, or worse, is penalized. The third type of error occurs when, given significant differences, there is a failure to account for or correctly identify the sources of differences. In this case, a personal predisposition toward teaching style may lead a user to attribute negative meanings to good ratings, or to misinterpret the results of an item as negative evidence when the item is actually irrelevant and there is no quantitative justification for such a decision (Franklin and Theall, 1990). Consequently, any of these errors could render an interpretation entirely invalid.

Additionally, there are differences in what specific instruments are intended to measure, how appropriate they are to different institutional settings, and how they should be used (e.g. teaching improvement or personnel decisions). For instance, it was noted that summary and global ratings, which are frequently used to make tenure and promotion decisions, were particularly elevated by instructor expressiveness. It was also found that lecture content had a sizable influence on student achievement, but only a negligible impact on student ratings. Findings such as these indicate that student instructional ratings should not be used to make decisions on faculty promotion and tenure, because they are based on lecturer characteristics (e.g. charismatic and enthusiastic) rather than student outcomes (Abrami *et al.*, 1982; Damron 1996).

Rater qualification error and defamation

Additionally, scholars have repeatedly indicated that students are not qualified to evaluate their lecturers. For instance, Adams (1997, p. 31) stated: "[Are] students, who are almost universally considered as lacking in critical thinking skills, often by the administrators who rely on student evaluations of faculty, able to critically evaluate their instructors? There is substantial evidence that they are not". Business literature clearly recommends that everyone who supplies data to be used in evaluation receive some kind of training. They may, however, be subject to legal challenge, because student ratings lack a certain degree of behavioural specificity (i.e. a five-point Likert scale) (Cascio and Bernardin, 1981). Further, if one is not qualified to perform a valid rating on another, the question of defamation exists. Normally, conversations between rating personnel are protected from defamation suits because of privilege (*Kasachkoff* v. *City of New York*, 485 N.Y.S.2d 992 (N.Y. App. 1985)). A privilege arises when both the speaker and the audience have a strong common interest. If, however, the party performing the rating is not qualified or does not possess a strong common interest, then privilege does not exist and defamation suits are plausible (*Colson* v. *Stieg*, 433 N.E.2d 246 (Ill. 1982)).

Case analyses

Each faculty member has at some point in their career questioned the reliability of SETE and the propriety of using the ratings for promotion and tenure decisions. The following cases illustrate the inconsistencies of

Quality Assurance in Education
Volume 11 · Number 1 · 2003 · 37-46

both student ratings and administrative interpretations of them. These examples are factual and drawn from personal experience of the authors.

Case 1. Reliable in meeting class
One of the authors has a habit of always arriving in the classroom five minutes earlier than the class is scheduled and he never missed a single class in the entire semester. However, the students rated him at 4.46 (on a five-point scale) in class A and 4.04 in class B, while the college average is 4.76. It is unclear how the lecturer might improve his performance toward the average.

Case 2. Available outside class
The same lecturer in Case 1 works over 90 hours per week on campus and is always available for student consultations. Yet, the students' evaluations of his performance on this variable were 4.28 in Class A and 4.17 in class B; these were lower than the college's average 4.55. Again, he was confused and has no idea how to improve his performance to be more available outside of class.

Case 3. Grading fair and reasonably
The same lecturer as in Cases 1 and 2 developed a grading system whereby students clearly understand an activity's grading rubric prior to completing the assignment, and the student papers and tests are graded without knowing the student's name. Students use pseudonyms when turning in their projects and report their real names only when picking up their graded activities. This system is designed to identify instructor expectations clearly and remove grader biases. Even so, students still rated him much lower than the college's average. For example, the college's average for this variable was 4.49, while his scores were 4.12 and 3.31 in class A and class B respectively. He does not know how he can be fairer and the students have failed to specifically outline complaints in the open-ended portion of the rating instrument.

Case 4. Prepared for class
Another lecturer spends considerable time developing her course syllabi. These syllabi list every class meeting time, the topics covered for each day, the reading and project assignments and due dates, and test dates. Additionally, the lecturer uses Power Point slides as visual aides to her lectures. She makes these slides available to her class on-line before each lecture. However, this lecturer received a sub-par rating for being prepared for class: she received a 4.64 and a 4.44, compared to a college average of 4.71, on a five-point scale. It is unclear how the lecturer might improve upon this average.

Case 5. Knowledge of subject
The above lecturer was rated 4.72 and 4.63 in her knowledge of the course. This was below the college average of 4.77 and was noted as an area of under-performance on her annual review. This lecturer never reuses a test, creates new projects for each semester, and only uses cases and examples drawn from the last 6 to 12 months. Clearly our question is not whether the lecturer is knowledgeable, but whether or not the students could accurately recognize knowledge of the subject and accurately define this perception on a five-point scale. Additionally, this case highlights the inappropriateness of describing 4.72 and 4.63 ratings (on a five-point scale) as under-performance.

Case 6. From excellent to average
The lecturer discussed in Cases 1, 2, and 3 is often publicly recognized by the dean as an example of someone whose students are exceeding academic expectations in terms of presentations, publications and standardized test scores. Ironically, however, the dean suggested during the annual performance evaluation that, based on SETE scores, the lecturer was below average in teaching. The lecturer discussed in Cases 4 and 5 was a finalist in a teaching excellence award. In this award, the college lecturers vote for the best examples of teaching excellence from the list of the top four lecturers as nominated by the students. On the year of her nomination, the dean noted that, although her student evaluations were slightly below college average, the students thought highly enough of her to earn her a finalist spot. Her teaching evaluation for that year was given as good/excellent. The very next year, with similar student evaluations and the criticisms noted above, the lecturer's overall teaching evaluation was given as satisfactory/good. This apparent inconsistency was brought to the dean's attention, who commented that although the student evaluation form may be flawed, it was all he currently had to use.

Quality Assurance in Education
Volume 11 · Number 1 · 2003 · 37-46

Case 7. Beneficial lab work
Two items on the evaluation instrument used at one college ask: is the lab work beneficial? And is the lab correlated with class? This instrument is the standard required of all lecturers at this liberal arts institution. Consequently, students in art, English, history, business and others are asked to evaluate the labs for courses that do not have labs. Naturally, the students may appropriately respond with "not applicable". However, as further evidence that students do not fully read or consider each item on the SETE forms, most courses without labs receive many evaluations from students on the appropriateness of these non-existent labs. In one course in particular, only 12 of 32 students marked "not applicable" on the evaluation form. The remainder rated the lecturer from 2 through 5. Naturally, the lecturer's average was significantly below that of the college.

Case 8. When is "good" good enough?
Upon considering the trials and tribulations of the above lecturers, it should be noted that the ratings are consistently above 4.0 on a five-point scale. This begs the question: when is "good" good enough? If a 4.5 is indicative of only satisfactory teaching, what does it take to get an "excellent" evaluation? If other factors are or should be taken into consideration, how then are they measured? If these "other" factors bear more weight than the students' evaluations, why then are the SETE still used? If the number is relevant only in comparison to others, this then creates another whole set of problems; particularly, if the college or university prides itself on hiring great teachers.

Case 9. The composition of the comparison group
This case demonstrates the inconsistencies found in the composition of the comparison groups. One lecturer received SETE scores of 4.10 and 4.24 for the two sections of a course he taught in the fall semester. The following spring semester, he taught the same two courses at the same college and received scores of 4.04 and 4.33. The college's average score for the fall semester was 3.99; for the spring it increased to 4.31. Compared longitudinally, the lecturer's scores were fairly consistent from one semester to the next. However, compared to the college averages, the lecturer's scores in the spring term were relatively lower than that of the fall term. Did the whole college's teaching effectiveness increase through the use of SETE? Obviously, it did not and the differences can be attributed, in part, to the composition of the faculty used to generate the college averages. In the fall term, all faculty members were required to submit SETE; while in the spring, only non-tenured and adjunct faculty were evaluated. It is generally accepted at these colleges that adjunct lecturers are usually "easier" in terms of expectations for the students; therefore, they would tend to get higher scores. Additionally, given the impact that SETE have on their careers, there may be a tendency among non-tenured faculty to ensure that students are "satisfied" to ensure that they receive above average evaluations. These two factors tend to inflate the overall college average. For the tenured faculty, the SETE scores will have a limited effect on their academic career and therefore they are insulated from the pressures to barter their educational standards for better student evaluations. Consequently, when tenured faculty evaluations are included, the overall average decreases. Does this suggest that tenured, more experienced faculty members are poorer teachers?

Discussion

It is exceedingly difficult to design and implement a performance appraisal process that is accepted as fair and just by all subordinates. Further, numerous research studies over the past several decades have suggested that they may be doing more harm than good (Levinson, 1965; McGregor, 1972; Meyer *et al.*, 1965; Mohrman, 1989). For example, Milliman and McFadden (1997) noted that General Motors discovered that 90 percent of its people believed they were in the top 10 percent. How discouraging is it to be rated lower? The following comments raised by industry can be legitimately made by higher education:

- They tend to foster mediocrity and discourage risk taking. The lecturer mentioned in Cases 1-3 has retreated from his rigourous expectations in order to receive higher student ratings. Unfortunately, student achievement has also retreated.

Quality Assurance in Education
Volume 11 · Number 1 · 2003 · 37-46

- They focus on short-term and measurable results, thereby discouraging long-term planning or thinking and ignoring important behaviours that are more difficult to measure.
- They focus on the individual and therefore tend to discourage or destroy teamwork within and between departments.
- The process is detection-oriented rather than prevention-oriented.
- They are often unfair, since administrators frequently do not possess observational accuracy.
- They fail to distinguish between factors that are within the faculty members' control and system-determined factors that are beyond their control.

Deming (1986) strongly condemned the performance appraisal process because of this last point. In the spirit of Deming, many companies are replacing performance evaluation altogether with personal planning and development systems. Several companies have replaced their traditional performance reviews with a personnel development planning process in which managers meet with employees to set future expectations, identify training needs, provide coaching, and reward continuous improvement. If higher education is going to fully embrace total quality, it requires a closely monitored performance appraisal process that is oriented toward "best practices" and continuous improvement of quality.

Further, the Porter and Lawler (1968) expectancy model of motivation can be used to illustrate the flaws in a process that fails to reward rigour and outcomes in teaching. If employees do not perceive a high effort-reward probability, they will not apply their best efforts to the task. As such, their abilities will not be exercised to the fullest, and their perceptions of their role in the organisation will be either negative or confused. This motivational problem is exacerbated when one perceives that the administration hopes for one thing and rewards another. Additionally, this confusion of objectives tends to diminish the lecturer's sense of organisational equity and procedural justice.

The crux of this problem might lie in the attempt to use the appraisal form for too many functions. Business organisations typically use performance appraisals to provide feedback to employees, who can then recognize and build on their strengths and work on their weaknesses, to determine salary increases, to identify people for promotion, to identify individual and organisational training needs, and to deal with human resource legalities. Perhaps education administrators should only use SETEs to collect qualitative information for feedback and focus on objective measures of outcomes for the teaching portion of promotion and tenure decisions. Further, if feedback is the primary purpose of using SETEs, then it seems logical to evaluate faculty members every semester, regardless of rank or tenure.

Conclusion

Considerable truth can be found in the statement, "How one is evaluated determines how one performs". This can be dangerous. We pose the question of whether higher education is evaluating "popularity" or "outcomes". Lecturers that perceive performance appraisals as popularity contests affecting their career will treat their students as customers rather than products. Several qualitative and quantitative studies have clearly supported the notion that higher education rewards the "self-interested" instructor (Comm and Mathaisel, 1998; DeBerg and Wilson, 1990; Delucchi and Pelowski, 2000). Further, Haskell (1997) points out that the majority of scholars believe that SETE represents a serious and unrecognized infringement of academic freedom. We doubt that the customers of a business school's products will feel comfortable with the knowledge that lecturers have sacrificed rigour for popularity and self-preservation.

Since the early 1970s a substantial literature has developed about faculty evaluation. Two excellent books have been published in the last ten years. The first published was *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness* by Centra (1993). The second was *Assessing Faculty Work: Enhancing Individual and Institutional Performance* by Braskamp and Ory (1994). We would suggest that there is almost universal agreement that the purpose of faculty evaluation is to help faculty improve their performance. However, an examination of the systems – as used – indicates that the

primary purpose is almost always to make personnel decisions (Cashin, 1996). We understand the need to make personnel decisions and that SETEs must be part of that decision process. As such, we offer the following list of recommendations to improve the use of student evaluations:

- Use multiple sources of data. Do not use student ratings as the only measure of teaching effectiveness. They do not provide evidence in all areas relevant to teaching effectiveness (e.g. command of subject matter, appropriateness of course content and objectives). Other useful sources for teaching effectiveness measurement could be the instructor's teaching portfolio, students' achievements, and peer evaluations.
- Make the wording on SETE instruments more "achievement" oriented rather than "satisfaction" oriented. Add questions that concern how much the students learned from the course and remove questions such as how well the instructors know the subject matter; students are not knowledgeable enough to make precise judgments (e.g. between a "4" and a "5").
- Rate faculty members against a standard rather than to a comparison of college-wide norms. For instance, it is appropriate (when using a well anchored five-point Likert scale) to rate all those course evaluations over 4.5 as excellent, those between 3.5-4.49 as good, those between 2.50-3.49 as satisfactory, and those below 2.49 as unsatisfactory. It is inappropriate to rate lecturers as "satisfactory" when their ratings are above 3.5 but below a faculty average that is above the above the "satisfactory" level. Additionally, any comparison should be performed against similar courses (e.g. a business course to a business course but not a business course to a music course).
- Ensure that the data/measures are technically acceptable, i.e. are reliable and valid. Use small sample size data for training or awareness purposes only, i.e. not for promotion, tenure and pay decisions. The data should be statistically evaluated for the purpose of eliminating undocumented extreme rating/outliers.
- Require students to specifically comment on ratings less than satisfactory. This will provide an opportunity to assess the credibility of negative ratings. Crumbley (1995) suggests that one way to make summative student evaluations more reliable is to require students to sign their names (or social security number). Under the present anonymous system, instructors have no due process for false and libelous statements. We believe this idea has considerable merit under a system where evaluations are viewed as constructive criticism. It would certainly help a lecturer to better understand the criticism, if he or she knew who was making the comment.
- Train the evaluators to evaluate and the supervisors in giving feedback. If students are to be an integral part of the unit's evaluation system, train them to evaluate during a freshman seminar. For example, instructors could discuss the meaning of student rating items with the students and practice rating various case studies. Further, the use of untrained evaluators may be subject to a legal challenge (Malos, 1998). Administrators, on the other hand, need to be trained in giving constructive feedback to prevent a reduction in motivation. If work behaviours rather than outcomes are to be evaluated, administrators should take the opportunity to observe the ratee's performance.
- Ensure that the system is legal. This is a complex topic that may require the consideration of several attorneys and precedence within education. For generally accepted "best practices", Centra (1993) has a chapter on "legal considerations in faculty evaluation" and Branskamp and Ory (1994) have some pages on "legal principles". For a reference on general legal questions see Kaplin and Lee's (1995) *The Law of Higher Education*.
- Ensure that the system is flexible. Any system of faculty evaluation needs to be concerned about fairness, which often translates into a concern about comparability. Using the same evaluation system for everyone almost guarantees that it will be unfair to everyone. Therefore, each academic unit should describe and give examples of how the institution's evaluation system applies to the characteristics and circumstances of that unit and its faculty.

- Ensure that the system celebrates diversity. All too often institutions take the approach that there is only one best way to teach. Implicitly this embodies the notion that what is different is dangerous and therefore unacceptable. Frankly, faculty from different cultures may honestly and justifiably have different concepts of what is acceptable and effective teaching behaviour. As such, units need to examine their evaluation systems to ensure that faculty members of different cultures are not receiving lower ratings because they are different.

In conclusion, we endorse the notion that "no one has taught anything, unless someone has learned something". As such, we encourage those programs that evaluate lecturers based on outcomes to come forward as models. We recognize that the activity of teaching is essentially one of human interaction, and as such is inextricably tied to the student's perception of a lecturer's personality. An evaluation of teaching effectiveness, however, must be based on outcomes. Anything else is rubbish.

References

Abrami, P.C., d'Apollonia, S. and Cohen, P.A. (1990), "Validity of student ratings of instruction: what we know and what we do not", *Journal of Education Psychology*, Vol. 82 No. 2, pp. 219-31.

Abrami, P.C., Leventhal, L. and Perry, R.P. (1982), "Educational seduction", *Review of Educational Research*, Vol. 32, pp. 446-64.

Adams, J.V. (1997), "Student evaluations: the ratings game", *Inquiry*, Vol. 1 No. 2, pp. 10-16.

Aleamoni, L. (1987), "Student rating: myths versus research facts", *Journal of Personnel Evaluation in Education*, Vol. 1, pp. 111-9.

Aleamoni, L. (1989), "Typical faculty concerns about evaluation of teaching", in Aleamoni, L.M. (Ed.), *Techniques for Evaluating and Improving Instruction*, Jossey-Bass, San Francisco, CA.

Arreola, R.A. (1995), *Developing a Comprehensive Faculty Evaluation System*, Anker Publishing, Boston, MA.

Braskamp, L.A. and Ory, J.C. (1994), *Assessing Faculty Work: Enhancing Individual and Institutional Performance*, Jossey-Bass, San Francisco, CA.

Cascio, W.F. and Bernardin, H.J. (1981), "Implications of performance appraisal litigation for personnel decisions", *Personnel Psychology*, Vol. 34, pp. 211-26.

Cashin, W.E. (1989), "Defining and evaluating college teaching", IDEA Paper No. 21, Center for Faulty Evaluation and Development, Kansas State University, Manhattan, KS.

Cashin, W. (1990), "Students do rate different academic fields differently", in Theall, M. and Franklin J. (Eds), *Student Ratings Of Instruction: Issues For Improving Practice*, Jossey-Bass, San Francisco, CA.

Cashin, W.E. (1996), "Developing an effective faculty evaluation system", IDEA Paper No. 33, Center for Faulty Evaluation and Development, Kansas State University, Manhattan, KS.

Centra, J.A. (1993), *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness*, Jossey-Bass, San Francisco, CA.

Cohen, P.A. (1983), "Comment on a selective review of the validity of student ratings of teaching", *Journal of Higher Education*, Vol. 54, pp. 448-58.

Colson v. *Stieg*, 433 N.E.2d 246 (Ill. 1982).

Comm, C.L. and Mathaisel, D. (1998), "Evaluating teaching effectiveness in America's business schools: implications for service marketers", *Journal of Professional Services Marketing*, Vol. 16 No. 2, pp. 163-70.

Crumbley, D.L. (1995), "The dysfunctional atmosphere of higher education: games professors play", *Accounting Perspectives*, Spring, Vol. 1 No. 1, pp. 27-33.

Damron, J.C. (1996), "Instructor personality and the politics of the classroom", available at: www.mankato.msus.edu/dept/psych/Damron_politics.html (accessed May 2001).

DeBerg, C.L. and Wilson, J.R. (1990), "An empirical investigation of the potential confounding variables in student evaluation of teaching", *Journal of Accounting Education*, Vol. 6, pp. 37-62.

Delucchi, M. and Pelowski, S. (2000), "Liking or learning? The effect of instructor likeability on overall ratings of teaching ability", *Radical Pedagogy*, Vol. 2, pp. 1-15.

Deming, W.E. (1986), *Out of the Crisis*, MIT Center for Advanced Engineering Study, Cambridge, MA.

Dooris, M.J. (1997), "An analysis of the Penn State student rating of teaching effectiveness: a report presented to the University Faculty Senate of the Pennsylvania State University", available at: www.psu.edu/president/cqi/cqi/srte/analysis.html (accessed May 2001).

Dowell, D.A. and Neal, J.A. (1982), "A selective review of the validity of student ratings of teaching", *Journal of Higher Education*, Vol. 53, pp. 51-62.

Dowell, D.A. and Neal, J.A. (1983), "The validity and accuracy of student ratings of instruction: a reply to Peter A. Cohen", *Journal of Higher Education*, Vol. 54, pp. 459-63.

Emery, C. (2001), "Professors as customers", in Lamb *et al.* (Eds) *Great Ideas for Teaching Marketing*, 6th ed., South-Western College Publishing, New York, NY.

Emery, C.R. (1995), "Student evaluations of faculty performance", manuscript, Clemson University, Clemson, SC.

Emery, C., Kramer, T. and Tian, R. (2001), "Customers vs products: adopting an effective approach to business students", *Quality Assurance in Education*, Vol. 9 No. 2, pp. 110-15.

Erikson, S.C. (1983), "Private measures of good teaching", *Teaching of Psychology*, Vol. 10, pp. 133-6.

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*, Vol. 9, pp. 199-242.

Feldman, K.A. (1986), "The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: a review and synthesis", *Research in Higher Education*, Vol. 24, pp. 139-213.

Franklin, J. and Theall, M. (1990), "Communicating student ratings to decision makers: design for good practice", in Theall, M. and Franklin J. (Eds), *Student Ratings of Instruction: Issues For Improving Practice*, Jossey-Bass, San Francisco, CA.

Haskell, R.E. (1997), "Academic freedom, tenure, and student evaluation of faculty: galloping pulls in the 21 century", *Education Policy Analysis Archives*, Vol. 5 No. 6, pp. 36-9.

Jones, J. (1989), "Students' ratings of teacher personality and teaching competence", *Higher Education*, Vol. 18, pp. 551-8.

Kaplin, W.A. and Lee, B.A. (1995), *The Law of Higher Education: A Comprehensive Guide to Legal Implications of Administrative Decision Making*, 3rd ed., Jossey-Bass, San Francisco, CA.

Kasachkoff v. *City of New York*, 485 N.Y.S.2d 992 (N.Y. App. 1985).

Levinson, H. (1965), "Appraisal of what performance?", *Harvard Business Review*, January/February, pp. 35-42.

McCallum, L.W. (1884), "A meta-analysis of course evaluation data and its use in the tenure decision", *Research in Higher Education*, Vol. 21, pp. 150-8.

McGregor, D. (1972), "An uneasy look at performance appraisal", *Harvard Business Review*, September/October, pp. 19-27.

McKeachie, W. (1987), "Can evaluating instruction improve teaching?", in Aleamoni, L.M. (Ed.), *Techniques for Evaluating and Improving Instruction*, Jossey-Bass, Inc., San Francisco, CA.

Main, J. (1994), *Quality Wars*, New York, NY.

Malos, S.B. (1998), "Current legal issues in performance appraisal", in Smither, J.W. (Ed.), Jossey-Bass, San Francisco, CA, pp. 49-94.

Meyer, H.H., Kay, E. and French, J.R. (1965), "Split roles in performance appraisal", *Harvard Business Review*, January/February, pp. 28-37.

Milliman, J.F. and McFadden, F.R. (1997), "Toward changing performance appraisal to address TQM concerns: the 360-degree feedback process", *Quality Management Journal*, Vol. 4 No. 3, pp. 44-64.

Mohrman, A.M. (1989), *Deming Versus Performance Appraisal: Is There a Resolution?*, Center for Effective Organisations, University of Southern California, Los Angeles, CA..

Porter, L.W. and Lawler, E.E. (1968), *Managerial Attitudes and Performance*, Irwin Publishing, Burr Ridge, IL.

Rosenfeld, P. (1987), *Instructor's Manual to Accompany Scarr and Vander Zadens' Understanding Psychology*, 5th ed., Random House, New York, NY.

Seldin, P. (1993), "The use and abuse of student ratings of instruction", *The Chronicle of Higher Education*, 21 July, p. A-40.

Sproule, R. (2000), "Student evaluations of teaching: methodological critique of conventional practices", *Education Policy Analysis Archives*, Vol. 8 No. 50, pp. 125-42.

Theal, M. and Franklin, J. (1990), "Student ratings of instruction: issues for improving practice", in Theal, M. and Franklin, J. (Eds), *New Directions for Teaching and Learning No. 43*, Jossey-Bass, San Francisco, CA.