Student Evaluations of Teaching Effectiveness: Considerations for Ontario Universities

By Mary Kelly, Wilfrid Laurier University

(With thanks to all the Academic Colleagues who provided institution-specific details and to John Sivell, Brock University; Linda Garcia, University of Ottawa; Philip Dutton, University of Windsor; Stephanie Crosty, David Johnson and Christine Neill, Wilfrid Laurier University; and Leslie Sanders, York University, for their input and comments)

May 2012

COU No. 866
ISBN: 0-88799-479-2
ISSN: 1704-412X (print)
ISSN: 1704-4197 (online)

The Discussion Paper Series consists of short papers on academic issues prepared by Academic Colleagues. Although each paper is discussed by the Colleagues and by Council, the final version of the paper represents the opinions of the author or authors. The papers as finalized do not represent COU policy. They are a mechanism for investigating and opening discussion on matters of interest to the Colleagues and Council.

Introduction

Student evaluations of teaching effectiveness or course evaluations (SETs)[1] are used at most universities and colleges throughout the world, and many researchers have analyzed their use in the classroom. Marsh (1987) commented that SETs may be the most studied form of personnel evaluation. There is a well developed literature addressing the construction of the student evaluations of teaching effectiveness. Much research has already examined the reliability and validity of the instrument, potential bias of student ratings, student perceptions of evaluations and their motivation to participate. However, despite the extensive amount of research, there is still much debate among academics as to the value and efficacy of student evaluations of teaching effectiveness.

We briefly discuss these findings where relevant, and examine the possible impact of evaluations on the teaching professoriate in Ontario. The measurement of teaching effectiveness is a particularly relevant topic given the increased focus on teaching and learning in Ontario universities.

- Fundamentally, student evaluations of teaching serve three basic purposes. From a formative (or developmental) perspective, evaluations are used to provide feedback to faculty in order to help them to improve their teaching or alter course content.
- When structured primarily as learning evaluations, SETs have also been used as a way to measure student engagement and learning – students use these forms to reflect on their self-learning.
- Evaluations can also serve summative (or administrative) purposes. They are used in tenure, merit and promotion decisions, and teaching evaluations also provide information to students to assist in course selection.

As noted by Gravestock and Gregor-Greenleaf (2008), SETs are more likely to be used for summative and not formative purposes.

We start with an overview of student perceptions of SETs and the impact on their willingness to participate. Fundamentally instructors are concerned with the use of SETs for summative purposes because there are concerns about the legitimacy and manipulability of SET scores. Therefore, we next focus on factors that impact SET scores that may not be directly correlated with teaching effectiveness. Finally, we discuss each of the three purposes of evaluations. Our discussion of the existing literature is supplemented by a sample of evaluations provided from 13 Ontario universities, and ad hoc discussions with professors in several Ontario universities. In our discussions, we comment on the efficacy of SETs in achieving these formative and summative goals, as well as other methods that have been used to attain these goals. We next discuss the impact of SETs on the faculty classroom management. One topic that is not addressed in the literature is the use of SETs as a measure of institutional accountability for excellence in learning outcomes and teaching excellence. We note limitations with the use of SETs for this purpose, but we also note the dearth of indicators to assess teaching excellence. Finally, we conclude with some observations of, and recommendations for, the use of SETs in Ontario universities.

Student Participation and Perception of SETs

In most institutions, students are not required to fill out SETs, and there is little incentive for them to fill out teaching evaluations. In fact, one Ontario university has eliminated "mandatory" end-of-term SETs due, in part, to poor response rate.

Students may not fill out SETs because they feel the evaluations are not valuable. Brown (2008) examines student perceptions of evaluations. He notes that students feel SETs can potentially provide an accurate measure of teaching effectiveness, but that students do not feel that administrators and other students take the evaluations seriously. Although students respond that they personally fill out the SET honestly, they believe that some other students are likely to rate instructors based on the grade they receive or use the SETs to "get back" at instructors.

If students feel that evaluations are not used to improve teaching performance or provide feedback on teaching effectiveness, they are less likely to fill them out. This perception of the value of SETs changes as a student progresses through university. Chen and Hoshower (2003) note that upper-year students are more likely to dismiss the usefulness of SETs, and as such participation rates tend to be higher in first-year classes. Kherfi (2011) also found that after controlling for respondent characteristics, first-year students are more likely to fill out teaching evaluations.

One topic of interest to teaching faculty is the potential switch to electronic evaluation and the resulting impact on both student participation and the SET results. Electronic evaluation of teaching effectiveness has significant monetary savings and allows for quicker feedback for instructors. However, research is unanimous that the use of electronic evaluations decreases participation levels over the traditional pencil-and-paper SETs that are typically administered in the last week of class. Layne, DeCristoforo and McGinty (1999) found response rates to be almost 13 percentage points lower for online versus pencil-and-paper evaluations (60.6 per cent versus 47.8 per cent). As with pencil-and-paper evaluations, they also find that students with higher GPAs are more likely to fill out electronic evaluations.

Avery *et al.* (2006) also found response rates to be consistently lower for online evaluations and conjecture that the reduction in participation may be driven by either the lack of peer pressure that arises when pencil-and-paper evaluations are filled out in class, or a fear concerning the loss of anonymity in electronic evaluations. However, they find that on a question-by-question basis, there was either no statistical difference in average scores or web scores were statistically significantly higher.[2] Consistent with earlier research, they also find that students who anticipated higher grades were more likely to submit an electronic evaluation. More recently, Venette, Sellnow and McIntyre (2010) compared pencil-and-paper versus online evaluation and, consistent with earlier results, found no significant difference in overall SETs scores between pencil-and-paper and electronic evaluations.

However, students provided more descriptive detail in the online than the paper-and-pencil responses. Therefore, when administered effectively, collecting student rating messages online may be better than doing so in the traditional pencil-and-paper format.

Davidovitch and Soen (2011) found that greater participation does not change overall evaluation. To improve participation in SETs, the University Centre of Samaria made the completion of SETs a mandatory requirement for graduation. There were no statistically significant differences in evaluations of individual professors before and after students were required to complete evaluations.

There may be significant privacy concerns when evaluations are conducted in small classes. This obviously is especially problematic for graduate and senior undergraduate courses. Most universities

attempt to mitigate this by not performing evaluations for small classes (with the threshold of "small" being defined by the institution). This may also result in some faculty having no SET scores if they only teach upper year or graduate classes.

Regardless of class size, doctoral students may be unwilling to provide feedback that may be construed as negative, due to the role of faculty in placing students in academic positions.

Perhaps there is a relationship between the quality of SET questions and response rates – students should be more engaged when the questions asked are meaningful and relevant. However, to the best of our knowledge, there is no concrete research on this aspect of the topic.

It has been suggested that improving communications to students about the value of SETs is one way to increase participation rates. One Canadian university suggests that faculty can encourage participation by highlighting changes made to a course based on previous evaluations.[3] Anecdotally, offering prizes or other incentives does not appreciably increase participation rates.

Overall, the literature would suggest that students appreciate the opportunity to provide feedback on teaching effectiveness – if the results of SETs are used to improve teaching performance and assist faculty in improving courses. Furthermore, research suggests that the use of electronic SETs may not pose as many problems as anticipated. Although the participation rate may be lower, the SET score does not change appreciably between pencil-and-paper and electronic evaluations.

Factors that Influence SET Scores

Effective teaching is multifaceted in nature. Therefore, it is not surprising that many factors influence students' perceptions on effective teaching. The literature indicates that students value instructors who are organized, provide clear and prompt feedback, create a classroom environment conducive to learning, respect students, and demonstrate concern for students (Bélanger and Longden, 2009).[4] Instructors who are more effective should fundamentally earn higher SET scores than instructors who are less effective. Overall, the research suggests that properly designed SETs are valid measures of teaching effectiveness (see, for example, Greenwald, 1997).

Research shows that SET scores are also influenced by external factors, such as the length of the course, class size, or age and gender of the instructor, and other factors that have no clear link to actual teaching effectiveness, such as student perceptions of what their final grade will be.

The relationship between SET scores and student grades is one external factor that has received much attention in the literature. There is a large body of literature that finds a positive correlation between grades and SET scores (see for example Marsh and Roche, 2000; Olivares, 2001; Griffin, 2004). This positive relationship leaves many to conjecture that instructors may be lenient in grading or develop easy assessments to influence SET scores. The evidence for this conjecture is mixed – some research has found that SETs are biased by grading leniency or reduced workload, and other research finds that these factors have little effect on SET scores. The relationship between SET scores and reduced course expectations is likely the most persistent of all concerns with the use SETs.

However, it is also possible that the correlation between grades and SET scores arises from student behaviour – students who are interested in a subject and are motivated to work hard, typically give higher evaluations and may also earn better grades. Finally, this relationship may actually be driven by more effective teaching. It is possible that the higher SET scores arise from student reactions to the classroom environment, which is correlated with the instructor's ability to teach. In addition, a classroom environment that is conducive to learning might naturally result in higher grades.

Supporting these latter two arguments, Kherfi (2011) found that students who do well (or anticipate that they will do well) in a course are more likely to participate in SETs. Remedios and Lieberman (2008) found that student ratings are largely driven by how well students feel they have been taught and how well they have been engaged, and are less influenced by self-reported workload and grade received (as long as the grade received was not lower than expected). In addition, they found that students rated courses to be more interesting and enjoyable if the grade they achieved in the course exceeded their expectations, which may capture to some extent the student's involvement in the course. Beran and Violato (2009) found similar results – engaged and motivated students give higher SETs scores.

Griffin (2004) argues that it is not the actual grade that matters, but student perceptions. If students anticipate that they will get a lower grade than they feel they deserve, they will rate the instructor lower in every question on the SET.

The average score that students give also changes as students progress through university. Chen and Hoshower (2003) found that seniors with higher GPAs generally give higher appraisals of professors, but in general first-year students hold faculty in higher regard than upper-year students.

Other research finds that, on average, instructors who offer midterm evaluations of teaching effectiveness perform better on SETs. However, it cannot be said that these are better teachers, as research also shows that the offering of these evaluations influences student perceptions. Instructors who conduct midterm evaluations of teaching effectiveness are perceived to be both committed to teaching and to value their students (Brown, 2008).

One persistent concern is that instructors who entertain students get higher SET scores. This argument is supported by Naftulin, Ware and Donnelly's (1973) experiment that suggested individuals could be "seduced" by a charismatic speaker who entertained and appeared to be knowledgeable. Abrami, Leventhal and Perry (1982) performed a meta-analysis of existing literature on relationships between course content, instructor expressiveness, student learning and SET scores. Their research found that SET scores were more responsive to changes in instructor expressiveness than to the lecture content itself.

It has been shown that students typically define an effective teacher as one who is warm, extroverted and enthusiastic (Clayson and Sheffet, 2006; Basow, 2000; Best and Addison, 2000). However, even with respect to these characteristics, students judge faculty based on age and gender. In an experiment, Arbuckle and Williams (2003) found that young male teachers score higher on warmth and enthusiasm. SET scores are also correlated with the gender of both the instructor and students. Feldman (1998) and Centra and Gaubatz (2000) both found that students tend to rank instructors of their own gender slightly higher.

Linse (2003), in a summary of gender and teaching studies, notes that gender effects exist but they are not uniform or simple to untangle: female faculty need to prove their intelligence, whereas male faculty are assumed to be intelligent; problematic teaching behaviours are more likely to be excused in men; generally, female professors are ranked lower on competency than male faculty; and students are more likely to view successful male faculty as effective and successful female faculty as likeable. However, the correlation between expectations and SET scores is difficult to untangle.

Course and class characteristics impact SET scores. McKeachie (1979) found that electives tend to have higher ratings than required courses. Bedard and Kuhn (2008) conclusively state that class size matters – there is a large negative impact of class size on SET scores even when controlling for the individual

instructor. Using a fixed effects regression model, they found that as class sizes increase, SET scores are lower. This is important for Ontario universities now as class sizes are growing.

Parpala *et al.* (2011) found that there is variation by discipline in students' beliefs as to what constitutes 'good teaching'. Thus, using a common evaluation tool across all disciplines may not be the best practice. Beran and Violato (2005) found that lab courses rate higher than lectures or tutorials. While it is 0understandable that active hands-on learning would be preferred by many students, issues arise when these SET scores are used for summative purposes and to compare instructors across different faculties.

Evaluations as a Feedback for Teaching Performance

We will now examine the use of evaluations used to provide feedback to faculty who wish to improve or alter their teaching practices. Here, evaluations are primarily focused on the use of faculty to assess the classroom experience. This is a formative use of SETs – instructors elicit feedback for their own use, in order to improve their teaching. Irrespective of how they are actually used, many of the SETs in our sample from Ontario universities state that this is the purpose of the evaluation form - for faculty review and improvement. In one Ontario university, this is the only use of SETs, as the results are available only to the faculty member.

There are concerns with the use of SETs for this formative use. Kherfi (2011) observes that the use of SETs to provide input on teaching effectiveness will be limited due to selection bias. Engaged students are more likely to fill out SETs than unhappy students: "*the underrepresentation of students who complete the course and obtain low grades implies that instructors receive proportionally less feedback from those who might be the most affected by any learning shortcomings associated with the instructor or the course*" (p.26).

The timing of SETs is also problematic. Typically SETs are conducted at the end of the semester; therefore, instructors do not have the ability to make changes for the course in which they are being evaluated. Because of this, many universities encourage their faculty to conduct their own midterm evaluations. If faculty can make changes to their course based on these midterm evaluations, it can be argued that midterm evaluations would be preferable. As noted above, the use of midterm evaluation of teaching tends to lead to higher end-of-semester scores for faculty.

However, the human resource literature finds that evaluations tend to be more positive and less critical than they should be if there will be future interactions between the evaluator and the instructor. This tendency is tempered by the use of anonymous evaluations. This unwillingness to criticize the instructor is likely to be stronger in midterm evaluations of teaching effectiveness because the term is not finished. It also could be significant in smaller departments or specialized programs if students know or perceive that they will be taught by the instructor in the future.

Initial research suggests the use of electronic SETs may improve the quality of student responses concerning teaching effectiveness. Venette, Sellnow and McIntyre (2011) found that although there was no difference in the number of positive and negative open-ended responses and no significant differences in the underlying themes of these responses between pencil-and-paper and electronic evaluations, students offered more descriptive electronic evaluations. This would indicate that the usefulness of student insights might be improved both at midterm and end-of-term evaluations of teaching if students could offer comments electronically.

The value of SETs depends on the quality of the survey, and concerns with students evaluating faculty teaching practices have been noted in the literature. Some questions are more useful and better tested

than others. Bélanger and Longden (2010) argue that faculty members are the best to judge the knowledge of the instructor, the appropriateness of the class objectives and the grading standards. Students, on the other hand, are better judges of classroom atmosphere, pace of instruction, and clarity and organization of class material.

Seldin (1993) observes that students, because of their limited background and experience, should not evaluate the instructor's knowledge of the subject matter or the materials used in the classroom. In a sample of 26 evaluations collected from 14 different Ontario universities, 4 evaluations asked students to rate the instructor's knowledge, and 14 asked if the resources and materials used in the course were effective and relevant.

Green *et al.* (1998) in a review of SETs used by accounting education departments in different institutions found that over 60 per cent of evaluations contain at least one question in which students did not have the background or experience to answer. Beran and Rokosh (2009) in their review at the University of Calgary found that most professors feel that the SET (at that institution) is too poorly designed to be useful to improve teaching practices. Spencer and Flyr (1992) in a faculty survey found than less than one-third of respondents used SET results to make meaningful changes to their teaching practices. Subsequent researchers argue that this is because faculty don't know how to meaningfully interpret statistics (Marsh, 2007). This rationale is suspect – it would imply that, for example, all statisticians because they can interpret statistics would be able to improve teaching performance (and following from that logic, they should have very high SET scores). We would argue that, in general, the survey questions asked do not provide adequate guidance to make meaningful changes to teaching practices.

In our sample of SETs from Ontario universities, six out of 14 institutions did not provide the opportunity for students to comment or provide written feedback to instructors, which would further reduce the usefulness of this form of evaluation.

Fundamentally, there are other ways to improve teaching skills than through the use of SETs. Most SETs are not a sufficient tool to improve teaching, as the questions are not specific enough. However, all universities have teaching support services/educational development groups to assist faculty members, although the use of their services is not compulsory. Typical supports are both internal and external conferences and workshops opportunities, mentoring and peer consultation, orientation and seminars for new faculty, and disseminating information on teaching innovation.

Wasley (2007) notes there are limitations to the use of faculty assisting other faculty in improving teaching skills. In commenting on Brigham Young University's program that has paid student observers providing feedback, she notes that student observers are less threatening to faculty members and provide a different perspective based on their experiences as students. In this service offered to faculty, student observers also interview other students in the course.

Evaluation as a Measure of Self-Learning

While the primary stated purpose of the sample evaluations examined were for "faculty review" or to "improve teaching", some evaluations asked for students to evaluate their learning in the course and others asked students to rate the value received in the course. Some universities are refocusing the role of evaluations to measure learning and not teaching. This is perhaps not surprising, given the recent focus on learning and engagement within the Ontario postsecondary education sector.

The questions asked on the SET impacts the students' perception of their roles in the learning process. Parpala *et al.* (2011) examine teaching in three different disciplines. They find that the clarity of

information and the teacher's efforts to make students understand are both important factors of good teaching from the students' point of view. Students are less likely to consider their own active role in the learning process. They contend that the SETs that focus on the role of teacher, influence students to minimize their active role in the learning process.

Industrial psychologists have examined the evaluation of training programs and self-reported learning. Kirkpatrick's seminal work on the evaluation of training programs (1959, 1960) notes that these evaluations do not capture learning *per se* (that is better captured by exam performance, for example), but instead student's reaction as to what the trainee thought of a particular program. Programs which are rated highly by participants may provide an appropriate atmosphere for learning, but these programs do not necessarily lead to high levels of learning. A welcoming and collegial atmosphere may be necessary for learning, but it is not sufficient. Goldstein and Ford (2002, p. 153) comment: "*It is important to realize that reaction measures may not be related to learning and eventual performance on the job. It is entirely possible for participants to enjoy the training but not produce the behaviour that is the objective of the instruction*."

In comparing our sample of SET questions to a list of survey questions assessing reaction measures (as provided by Grove and Ostroff (1991) and summarized by Goldstein and Ford (2002)), it is readily apparent that SET questions, in general, are not measures of self-learning but reactive measures of enjoyment. These five statements are given below and the number in brackets beside each question is the number of surveys in our sample of SETs that asked the same or similar question:

- the objectives of this program were clear (7)
- the instructor was helpful and contributed to the learning experience (17)
- there was an appropriate balance between lecture, participant involvement and exercise in the program (10)[5]
- the topics covered in this program were relevant to the things I do on my job (4)
- overall how you would rate this program (14 rated course and 20 rated instructor)

Thus, these questions, which are used by universities to rate instructor performance or student learning, appear to be measuring student reactions instead. That is: they measure how a student enjoyed the learning environment, not whether or not they have learned the material.

And, finally, perhaps students may not know what they need to know. Guevara and Stewart (2011) note that alumni have different course expectations than undergraduate students, and, in particular, material that has career relevance and learning is a stronger determinant of course satisfaction for alumni than undergraduates.

Evaluations of Teaching Performance

For teaching faculty, the summative role of evaluations is the most contentious issue.[6] Evaluations are used to measure teaching effectiveness for use in tenure and promotion decisions, and in assessing merit. Given the importance of these evaluations in tenure and promotion decisions, faculty are understandably concerned about the legitimacy (and, perhaps, manipulability) of the results. However, overall the research finds that properly designed SETs are both valid and reliable.

Gravestock and Gregor-Greenleaf, in a presentation to COU Academic Colleagues, are concerned that because consistency is important in ensuring validity, results of evaluations collected electronically outside of class time should not be compared to results of evaluations conducted in class. This concern may be unfounded, as Avery *et al.* (2006) note that SET scores should not be adversely affected by

switching to electronic scores. Nonetheless, lower response rates may lead to greater variability in SET results, which makes it difficult to track the change in a faculty's teaching effectiveness over time.

As discussed above, student perceptions of SETs and their usefulness vary across class size, gender and year of student, and how the evaluation is administered. Differences in perceptions may lead to differences in SET scores. Therefore, when SETs are used for summative purposes, there may be unintended biases against some faculty members. The SET should not form the only basis for the evaluation of faculty teaching competence. Alternatively, McPherson, Jewell and Kim (2009) provide a suggestion on how SET scores could be adjusted for these extrinsic factors outside the control of the instructor.

How SET scores are used in the tenure and promotion process also varies between and across universities. Several Ontario universities use the average score of all SET questions to influence decisions about tenure and promotion. Statistically, this averaging of averages typically results in very few outlier scores. Thus, it becomes difficult to distinguish across levels of teaching effectiveness except for the most exceptional and poor teachers. One university states that only this aggregate score, and no individual SET question score or student input, can be used by the university in its assessment of teaching performance. Many universities report the score from the summative question on the SET, as well as the average score of all SET questions, as an input to assessment of teaching performance. Some universities have two summative questions – one on the quality of the instructor, and one on the quality of the course. The impact and importance of these evaluations varies according to the instructor's employment. SET scores may be the main input for rehiring of sessional and part-time contract instructors, and they are typically more important to teaching-only faculty than regular faculty members. Since SET scores are a part of the tenure and promotion process, these evaluations are more important for tenure track than tenured faculty. Some universities in Ontario solicit input from graduate and undergraduate students on teaching effectiveness for tenure and promotion decisions, but most do not and, indeed, many collective agreements do not allow for written input on teaching effectiveness from students.

A further concern about the use of SET scores in personnel decisions has been raised by Hoyt and Pallett (1999). SET scores are often used to make comparisons between instructors in different departments or faculties. This is only valid if institutions have standardized survey questions across different faculties and divisions. Even then, there is disagreement in the literature about whether or not SETs can be used to compare faculty members. At most, as noted by Gravestock and Gregor-Greenleaf (2008), comparisons can only be reliably made across instructors in the same discipline.

A teaching dossier or portfolio, which typically contains teaching material (course outlines, evaluations and assignments), a summary of teaching accomplishments and a statement of teaching philosophy, is also required in the tenure process by many Ontario universities. As noted by the collective agreement at one Ontario university, this dossier "permits faculty to provide a context for student evaluations". However, even with the dossier, SET scores are arguably still the most important measure of teaching effectiveness.

Impact of Evaluations on Faculty Performance

There have been a few extensive studies on faculty perception of SETS (Nasser and Fresko, 2002), as most studies are typically localized within one university, or even one faculty within a university. Thus, even the published research on faculty perceptions is largely anecdotal in nature.

Beran and Rokosh (2009) summarize the literature on faculty perceptions of SETs, and note that there is very little consensus - many instructors view SETs to be valuable and many do not. Their research, which included a sample of over 300 faculty members at a single Canadian university, found that two-thirds of the respondents expressed a negative view of the SET used at their institution. Concerns at this institution (and we would argue that these concerns are relevant to Ontario teaching faculty as well) arose primarily from the quality and legitimacy of SET scores. Are SET scores biased by factors outside of the faculty member's control? Do faculty manipulate (or manage) the classroom environment to improve SET scores in ways that may worsen the teaching and learning experience?

Ackerman, Gross and Vigernon (2009) undertook in-depth interviews with faculty members to categorize their perception of SETs. Every instructor interviewed expressed concern that SET scores are impacted by grades or grade expectations, and half of instructors felt pressured to please students rather than provide an honest assessment of performance. Nasser and Fresko (2002), in their survey of 100 faculty members at a teacher's college, found that faculty believed that demanding instructors and challenging courses get lower SET scores but, overall, good instructors in general earn higher scores. These views have been expressed previously in the literature. Haskell (1997) hypothesized that instructors may feel unnecessarily constrained in providing written feedback to students because instructors worry that students who disagree with written feedback will provide poorer evaluations. Penny (2003) and DeNisi and Kluger (2000) suggest that faculty are less likely to try innovative teaching approaches because they fear poor evaluations. Beran and Rokosh (2009) found that some instructors are unwilling to stray from the course outline, even if that would produce a richer pedagogical experience, because this will produce lower SET scores. In fact, over half of Ontario SETs collected, asked students whether the course was consistent with the course outline.

One area not fully explored in the literature is the impact of evaluations on faculty decisions regarding classroom management that go beyond pedagogical considerations. The empirical literature on the ability of faculty to influence SET scores by giving high grades, or reducing work load, is mixed. However, many faculty perceive that it is possible the SET scores can be managed, either intentionally or unintentionally. This perception may influence classroom management decisions.

For example, one often cited piece of advice offered to new instructors is to never return student assignments/grades on the day that SETs are done, unless grades exceed expectations. And some faculty, in deciding both beforehand what guidelines are enacted at the beginning of term and what rules are enforced during the term, consider the impact that these decisions will have on teaching evaluations. Classroom management decisions might include:

- Will there be flexibility in deadlines and will late assignments be allowed?
- Will extra credit work be given if students are performing poorly?
- Will all incidents of academic misconduct be rigorously pursued?
- Will treats or snacks be provided to students, especially in smaller classes? And if so, when is the right time to offer them (before any SET is administered, or after SET is administered)?
- What sort of classroom discipline will be imposed (attendance and deportment)?

The concern here is with decisions that are made to manipulate SET scores, but that do not improve the teaching or learning experience, or that are not pedagogically appropriate.

Consistent with themes found in the literature, faculty may be less strict than they might otherwise wish to be, because they do not want students to give them poor evaluations. Added to this is the pressure to increase student retention and improve the student experience, as well as the pressure from increasing class sizes and scarcity of resources. These effects also influence instructors. Fundamentally, even if

there is no statistical relationship between SETs scores and instructor leniency, the fact that instructors buy in to this "belief" will have a negative impact on the quality of the course taught.

It seems plausible that this concern is greater for faculty members whose positions or future job opportunities are more dependent on good teaching evaluations – graduate students, contract academic staff and tenure track faculty. As Ontario universities rely more and more on contract staff to teach, these impacts are more pronounced.

University Accountability and SETs

Universities in Ontario are under increasing pressure from students, the families of students and, indeed, the general population to achieve excellence in teaching. As noted by the Higher Education Quality Council of Ontario (HECQO), core indicators of teaching excellence focus on inputs, such as class size, or institutional instructor-student ratios, and not on output measures of student learning.

Despite their limitations, SETs are one of the few quantitative measures of teaching performance even if, at most, what is measured are student perceptions. Yet, in a survey of 14 Canadian universities, Gravenstock and Gregor-Greenleaf (2008) found that only six of the universities examined report some measure of SET data to the university community. In the absence of any other indicator, how can Ontario universities measure their commitment to improving teaching performance?

What is more relevant than teaching excellence is successful learning outcomes. Research has found positive correlations between SET scores and amount learned (Theal and Franklin, 2001). But Berk (2005, p 56.) notes that "*key characteristics of students, such as ability, attitude, motivation, age, gender, and maturation, and of the institution, such as class size, classroom facilities, available technology and learning resources, and school climate, can affect student performance irrespective of what an instructor does in the classroom.*"

Rationally, the only way to directly test learning outcomes is to test the students themselves. Although beyond the scope of this paper, organizations in the United States, such as the National Institute for Learning Outcomes Assessment and the Voluntary System of Accountability Program, have developed frameworks and best practices for both the testing of learning outcomes and dissemination of the institutional accountability measures.[7]

Conclusions and Final Thoughts

There are some unintended but positive impacts from SETs that we have not yet mentioned. Even if faculty are concerned that students may not provide the best evaluations of teaching performance, the mere presence of SETs underscores the responsibility that instructors have to students. As noted by Beran and Rokosh (2009), SET "introduces some measure of responsibility towards one's students." (p. 507).

Our observations and the above literature suggest that the design of SETs matter. Indeed, some universities are currently revising their institutional policies with respect to the use of SETs. However, many universities are intent on recreating the wheel, as there are existing surveys of teaching evaluation that are well tested for validity and reliability. One example is the Students' Evaluation of Educational Quality or SEEQ, which is used globally at several institutions.[8]

From a formative perspective, a simple evaluation on its own is not worthwhile – evaluation is only valuable when it leads to improvements in teaching. If a SET is deemed to be essential to this process,

then the questions on the SET must actually evaluate inputs to teaching effectiveness. It may be the case that, at most, SETs evaluate whether or not the classroom experience is conducive to learning.

The survey of the literature reveals that there are few best practices for the administration and use of SETs. We briefly summarize our key findings and observations that would support best practices in evaluation and accountability.

- All universities should ensure that their surveys questions are valid and reliable. Using an accepted survey, such as the Students' Evaluation of Educational Quality (SEEQ) is one way to ensure this.
- At most, SETs should be only one component of faculty evaluation for personnel decisions, such as tenure, promotion and merit.
- From a summative perspective, properly designed SETs may be valid and reliable, but even then there are uncontrollable factors (such as class size, age and gender of instructor) that impact SET scores. Thus, using SETs as the sole basis to measure teaching effectiveness for promotion and tenure decisions is not appropriate.
- SET scores cannot be used to compare faculty members across different faculties or departments in a university if the underlying survey instrument (or the manner in which SETs are administered) varies across faculties.
- Measuring teaching effectiveness requires input such as teaching dossiers and evaluation by faculty members. Faculty members are better at evaluating course content, whereas students are the best at evaluating what worked or did not work for their learning. Thus, Ackerman, Gross and Vigernon (2009) recommend that faculty observations should be used in conjunction with SETs to provide an accurate picture of teaching effectiveness. Berk (2005) provides 12 sources of evidence of teaching effectiveness – including a combination of student, alumni, employer and administrator rankings. Innovative programs that involve student input in improving university education, like the one at Brigham Young University (Wasley, 2007), should also be considered.
- Students are more likely to take SETs seriously if they see that their input matters. Faculty can show that student input matters by undertaking midterm evaluations of teaching effectiveness and then making changes within the course. Faculty members can point out to students how previous input has helped shaped the course. Gravestock and Gregor-Greenleaf (2008) suggest that providing students with (a subset) of previous SET scores for a course/instructor also improves the usefulness and appropriateness of student input.
- SETs do not improve teaching; institutional teaching support and programming improves teaching. Universities need to commit to incentives and structures for faculty members to strive for teaching excellence.
- Institutional accountability for teaching excellence and learning outcomes is important. Current SET scores are the only quantitative indicator collected by most universities that attempt to measure, at some level, teaching excellence. Most would agree that it is not appropriate to make SET scores publicly available. However, universities should strive to create publicly available institutional-level performance measures that report on both teaching quality and learning achievements.

Finally, human resource specialists would argue that the use of a single tool for many purposes (measuring student learning, and measuring and improving teaching effectiveness) is not ideal, and it is better to have a tool designed for each objective. Although SETs are entrenched in our university system, there continues to be a growing concern for student learning accompanied by a lack of support for

university professors to teach effectively. It is, perhaps, important for universities to rethink how these are done and, more importantly, what purpose such data serves in the university system.

References

Abrami, P.C., L. Leventhal and R.P Perry (1982) Educational Seduction. *Review of Educational Research* **52**(3): 446-464.

Agbetsiafa, D. (2010). Evaluating Effective Teaching in College Level Economics Using Student of Instruction: A Factor Analytic Approach. *Journal of College Teaching and Learning* **7**(5):57-66.

Ackerman, D., B.L. Gross and F. Vigneron (2009). Peer Observation Reports and Student Evaluations of Teaching: Who Are the Experts? *Alberta Journal of Educational Research* **55**(1): 18-39.

Arbuckle, J. and B.D. Williams (2003) Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations. *Sex Roles* **49**: 509-516.

Avery, R.J., W.K. Bryant, A.M.H. Kang, and D. Bell. (2006) Electronic Course Evaluations: Does an Online Delivery System Influence Student Evaluations? *Journal of Economic Education* **37**:21–37.

Basow, S.A. (2000). Best and Worst Professors: Gender Patterns in Students' Choices. *Sex Roles* **34**: 407–417.

Bedard, K. and P. Kuhn (2008). Where Class Size Really Matters: Class Size and Student Ratings of Instructor Effectiveness. *Economics of Education Review* **27**: 253-265.

Bélanger C.H. and B. Longden (2009). The Effective Teacher's Characteristics as Perceived by Students. *Tertiary Education and Management* **15**(4): 323-340.

Beran, T.N., and J.L. Rokosh (2009). The Consequential Validity of Student Ratings: What Do Instructors Really Think? *Alberta Journal of Educational Research* **55**(4): 497-511.

Beran, T., and C. Violato (2005). Ratings Of Teacher Instruction: How Much Do Student And Course Characteristics Really Matter? *Assessment and Evaluation in Higher Education* **306**: 593-601.

Beran, T., and C. Violato (2009). Student Ratings of Teaching Effectiveness: Student Engagement and Course Characteristics. *The Canadian Journal of Higher Education,* **39**:1-13.

Berk, R.A. (2005). Survey of 12 Strategies to Measure Teaching Effectiveness. *International Journal of Teaching and Learning in Higher Education* **17**(1): 48-62.

Best, J.B. and W.E. Addison (2000). A Preliminary Study of Perceived Warmth of Professor and Student Evaluations. *Teaching of Psychology* **27**: 60–62.

Brown, M. J. (2008). Student Perceptions of Teaching Evaluations. *Journal of Instructional Psychology* **35**(2):177-181.

Centra, J. A., & Gaubatz, N. B. (2000). Is There Gender Bias in Student Evaluations of Teaching? *Journal of Higher Education* **71**(1):17-33.

Chen, Y. and L.B. Hoshower (2003). Student Evaluation of Teaching Effectiveness: An Assessment of Student Perception and Motivation. *Assessment & Evaluation in Higher Education* **28**(1):71-88.

Clayson ,D.E. and M.J. Sheffet (2006). Personality and the Student Evaluation of Teaching. *Journal of Marketing Education* **28**(2): 149-160.

Davidovitch, N., & Soen, D. (2011). Student Surveys and Their Applications in Promoting Academic Quality in Higher Education. *Journal of College Teaching and Learning* **8**(6):31-46.

DeNisi, A.S., & Kluger, A.N. (2000). Feedback Effectiveness: Can 360-Degree Appraisals Be Improved? *Academy of Management Executive* **14**(3): 129-139.

Feldman, K. A. (1977) Consistency and Variability among College Students in Their Ratings Among Courses: A Review and Analysis. *Research in Higher Education* **6**(3): 223–274.

Goldstein, I.L. and J.K. Ford (2002). *Training in Organizations: Needs Assessment, Development, and Evaluation (4th ed.)*. Wadsworth/Thomson Learning: Belmont, CA.

Gravestock, P. and E. Gregor-Greenleaf (2008). *Student Course Evaluations: Research, Models and Trends.* Higher Education Quality Council of Ontario: Toronto, ON.

Green, B.P., T.G. Claeron and B.P. Reider (1998). A Content Analysis of Teaching Evaluation Instruments Used in Accounting Departments. *Issues in Accounting Education* **12**(1): 15-30.

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

Greenwald, A.G., and G.M. Gillmore (1997). No Pain, No Gain? The Importance of Measuring Course Workload in Student Ratings of Instruction. *Journal of Educational Psychology* **89**(4): 743-751.

Griffin, B.W. (2004). Grading Leniency, Grade Discrepancy, and Student Ratings of Instruction. *Contemporary Educational Psychology* **29**(4): 410-425.

Grove, D.A. and C. Ostroff (1991). Program Evaluation. In *Developing Human Resources*. (eds: K. Wexley and J. Hinrichs). BNA Books: Washington, DC

Guevara, C. and S. Stewart (2011). Do Student Evaluations Match Alumni Expectations? *Managerial Finance* **37**: 610-623

Haskell, R.E. (1997). Academic Freedom, Tenure, and Student Evaluation of Faculty: Galloping Polls in the 21st Century. *Education Policy Analysis Archives* **5**(6): 1-35.

Hoyt, D.P. and W.H. Pallett (1999). *Appraising Teaching Effectiveness: Beyond Student Ratings* (IDEA Paper No.36). Kansas State University Center for Faculty Evaluation and Development:

Manhattan, KS., P. Isely and H. Singh (2005). Do Higher Grades Lead to Favorable Student Evaluations? *Journal of Economic Education* **36**(1): 29-42.

Kherfi, S. (2011). Whose Opinion Is It Anyway? Determinants of Participation in Student Evaluation of Teaching. *Journal of Economic Education* **42**(1): 19-30.

Kirkpatrick, D.L. (1959). Techniques for Evaluating Training Programs. *Journal of the American Society of Training Directors* **13**: 3-9, 21-26.

Kirkpatrick, D.L. (1960). Techniques for Evaluating Training Programs. *Journal of the American Society of Training Directors* **14**: 13-18, 28-32.

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic Versus Traditional Student Ratings of Instruction. *Research in Higher Education* **40**(2): 221-232.

Linse, A.R. (2003). Student Ratings of Women Faculty: Data and Strategies. University of Washington Working Paper. Available on line at
http://advance.washington.edu/apps/resources/docs/20030513-student_ratings_ds.pdf.

Marsh, H. W. (1982). SEEQ: A Reliable, Valid, and Useful Instrument for Collecting Students' Evaluations of University Teaching. *British Journal of Educational Psychology* **52**: 77-95.

Marsh, H.W. (1987). Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. *International Journal of Education Research* **11**(3): 253-388.

Marsh, H.W. (2007). Do University Teachers Become More Effective with Experience? A Multilevel Growth Model of Students' Evaluations of Teaching Over 13 Years. *Journal of Educational Psychology* **99**(4): 775-790.

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

McKeachie, W.J. (1979). Student Ratings of Faculty: A Reprise. *Academe* **62**: 384-397.

McPherson, M. A., R. T. Jewell, and M. Kim. (2009). What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes*. Eastern Economic Journal* **35**: 37–51.

Mukherji, S., & N. Rustagi. (2008). Teaching Evaluations: Perceptions of Students and Faculty. *Journal of College Teaching & Learning* **5**(9): 45-54.

Naftulin, D.H., J.E. Ware, and F.A. Donnelly (1973). The Doctor Fox Lecture: A Paradigm of Educational Seduction. *Journal of Medical Educatio*n **48:** 630-635.

Nasser F. and B. Fresko (2002). Faculty Views of Student Evaluation of College Teaching. *Assessment & Evaluation in Higher Education* **27**(2): 187- 198.

Olivares, O.J. (2001). Student Interest, Grading Leniency, and Teacher Ratings: A Conceptual Analysis. *Contemporary Educational Psychology* **26**: 382–399.

Parpala, A., S. Lindblom-Yläänne and H. Rytköönen (2011): Students Conceptions of Good Teaching in Three Different Disciplines. *Assessment & Evaluation in Higher Education* **36**(5): 549-563.

Penny, AR. (2003). Changing the Agenda For Research Into Students' Views About University Teaching: Four Shortcomings of SRT Research. *Teaching in Higher Education* **8**: 399-411.

Remedios, R., and D. A. Lieberman. (2008). I Liked Your Course Because You Taught Me Well: The Influence of Grades, Workload, Expectations and Goals on Students' Evaluations of Teaching. *British Educational Research Journal* **34**: 91-115

Seldon, P. (1993). The Use and Abuse of Student Ratings of Professors. *The Chronicle of Higher Education* **40**(1): 40.

Spencer, P.A. and M.L. Flyr (1992). *The Formal Evaluation as an Impetus to Classromm Change: Myth or Reality.* ERIC Report: ED34653.

Venette, S., D. Sellnow and K McIntyre (2010). Charting New Territory: Assessing the Online Frontier of Student Ratings of Instruction. *Assessment & Evaluation in Higher Education* **35**(1): 97-111

Wasley, P. (2007). How Am I Doing? *Chronicle of Higher Education* **54**(9): 10.

[1] These evaluations have different names at different universities and in the literature. For consistency purposes, we adopt the name "student evaluation of teaching" or SET in this document.
[2] However, this result should be interpreted with caution. Among this sample was one professor who, when teaching two sections of a course, used paper evaluation twice in one section and electronic evaluation in another. In one semester, his evaluations were higher in the paper format and, in the next semester, they were higher in the electronic format.
[3] http://www.mcgill.ca/tls/teaching/course-evaluations/completion
[4] This is a European study. Results from North American studies are similar. However, one might suspect that actions represented by "treating students respectfully" or "creating a classroom conducive to learning" would be different in different cultures.
[5] The phrasing of the SET questions typically asked if the articles, discussions and projects were effective, or if the instructor used a variety of teaching styles.
[6] Another summative role is providing summary teaching evaluation scores to students to assist in course selection. We conjecture that many faculty might prefer this to ratemyprofessors.com, which tends to attract more extreme comments.
[7] See, for example, http://www.learningoutcomeassessment.org/ and http://www.voluntarysystem.org/index.cfm
[8] See, for example, Marsh (1982), Marsh (1987), http://www.mta.ca/pctc/TONI_SEEQ/SEEQ_long.pdf and http://www.schreyerinstitute.psu.edu/Tools/SEEQ/Analyze/.