Communicating the outcomes of student evaluations of teaching: one-size-fits-all?

Marcia Devlin
mdevlin@deakin.edu.au

Abstract: Student evaluations of teaching are increasingly used to measure the teaching of individual academics in Australian higher education. The outcomes of these evaluations are variably made available to the individual academics themselves, to university management and to the public. However, communicating evaluation outcomes to each of these audiences assumes a different purpose and necessitates different objectives, foci and methodology. The need for these differences is sometimes forgotten with a ‘one-size-fits-all’ approach in Australian higher education. This paper examines these differences and discusses some of the issues surrounding the communication of the outcomes of student evaluations of teaching.

Student ratings of teaching have steadily taken precedence in teacher evaluation systems in North America and Australia and are growing in popularity in Asia and Europe (Theall and Franklin, 2000). In the UK, since the 2002 Higher Education Funding Council for England (the so-called Cooke Committee) specified quality information requirements for HE quality assurance and this was reinforced by the 2003 UK Government White Paper on the future of higher education, student feedback has become an important element in the quality process (Harvey, 2003). As Harvey points out, the advent of quality assurance in teaching has highlighted the expectation that universities will regularly investigate to what extent aims in relation to teaching and learning are being realised and standards are being maintained. In addition, evidence that feedback has been sought and responded to is increasingly required as part of confirmation and promotion procedures (Hounsell, 1999).

In Australia, the proposed Learning and Teaching Performance Fund, as part of the policy outlined in Our Universities: Backing Australia’s Future (DEST, 2003) specifies that funding allocations will be determined once institutions meet specific teaching related requirements. These include “…probation and promotion practices and policies that include effectiveness as a teacher as a criterion for those academics with a teaching load…” and “…systematic student evaluation of teaching and subjects that informs probation and promotion decisions…” for these academics. Further, student evaluation results will have to be made public on an institution’s website.

The purposes of SETs

In addition to providing data for research on teaching, the outcomes of student evaluations of teaching (SETs) can be used for three main purposes. The first is to provide diagnostic feedback to teaching staff about the effectiveness of their teaching with the intention that this will be useful in improving their teaching. The second is to provide a measurement of teaching effectiveness to be used in confirmation, promotion and other administrative and personnel decision-making. The third is to provide information for prospective and continuing students to use in selecting subjects/teachers (Marsh and Dunkin, 1997).

The recent proposed changes to Australian higher education indicate that the second and third of the three reasons outlined above may achieve prominence in coming years, perhaps, but hopefully not, at the expense of the first. In any case, the proposed DEST requirements hint at confusion about the different purposes of collecting SETs. Table 1 summarises and the discussion below examines the differences in objectives, foci and the communication of findings when the purposes of collecting and communicating the results of SETs vary.

<table>
<thead>
<tr>
<th>Intended purpose</th>
<th>Objective(s)</th>
<th>Focus</th>
<th>Method</th>
<th>Communication of SET data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Feedback to teaching academics</td>
<td>- Formative</td>
<td>- Areas in need of enhancement</td>
<td>Collect SETs; target areas in need; collect SETs again</td>
<td>- Private - Between staff member and peer or academic developer</td>
</tr>
<tr>
<td></td>
<td>- Diagnostic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Measurement of teaching for personnel decisions</td>
<td>- Summative</td>
<td>- Overall competence</td>
<td>Collect quantitative and qualitative SET data over time</td>
<td>- Confidential, internal - Between staff member and supervisor/committee</td>
</tr>
<tr>
<td></td>
<td>- To inform performance appraisal</td>
<td>- Areas of excellence</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Areas that have improved</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Information for future teaching/subject decisions</td>
<td>- Summative</td>
<td>- Overall performance</td>
<td>Collect SET data at series of end points (after diagnostic and personnel)</td>
<td>- Public - Anyone with access to website can view data</td>
</tr>
<tr>
<td></td>
<td>- To inform public</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1: Summary of major aspects of three types of student evaluation of teaching**

**Communicating the outcomes of SETs for diagnostic and developmental purposes**

When SET outcomes are to be used to communicate with staff about their teaching, they should be part of a formative process for that staff member. The focus should be on areas of teaching in which students believe enhancement or improvement is most needed. Ideally, SETs for this purpose are designed to gather data to inform teaching-related goal setting. Teaching strengths are acknowledged but perceived teaching weaknesses are examined and used as the basis for intervention that might include an individual program of development, the adoption of particular strategies, mentoring, workshops, reading and/or reflection, with or without the assistance and input of a peer or an academic developer.

It is important to note that it is essential that the process in these cases is embarked on and maintained voluntarily. Successfully identifying, documenting and specifically addressing one’s professional weaknesses requires a process where there is confidentiality, trust and collaboration between the staff member whose teaching is being evaluated and any staff involved in conducting, interpreting or communicating the SET results to that staff member. Ownership of the process by the teacher who is at the centre of the evaluation is also central if genuine attempts at improvement are expected. For obvious reasons then, the results of SETs gathered to inform development, enhancement and improvement should be communicated only to the staff member on whose teaching student views are focused.

**Using SETs for judgemental, personnel purposes**

When SETs are to be used to make personnel decisions, the main foci should be overall competence and areas of teaching excellence. Aspects of teaching that have improved may also be highlighted by the staff member in a submission or application. The objective is to provide a summative measure of teaching. Ideally, qualitative and contextual information, including an academic’s philosophy of teaching, views on student learning and the like would be combined with the quantitative in order to provide a detailed picture of teaching development and competence as well as areas of excellence.
As they stand, the DEST proposals do not seem to allow for the compilation of this fuller picture. While the current proposal specifies the use of SETs to inform probation and promotion decisions, it does not specify any other source of relevant data that may equally inform such decisions.

One major difference between using SETs for diagnostic/development purposes and their use for judgemental/personnel decisions is that while the former is necessarily a voluntary process, the latter is likely to be mandatory. But where insufficient opportunity and resources have been provided for teaching staff to develop their teaching and insufficient rewards for teaching are available, the use of SETs for decision making about a teaching academic’s career progress may become problematic.

Academic staff are generally employed for their research expertise and most are not qualified teachers. Indeed, it is usually, but not always, the desire to pursue research in a particular field, and not a desire to teach, that is the primary factor that leads an academic to seek university employment. Teaching is viewed, by the majority of academics, as a necessary duty rather than the focus of one’s position and teaching is viewed, by the majority of universities, as ‘second fiddle’ to research in terms of career advancement. It is therefore simply not fair to expect or mandate that academic staff focus on excelling in teaching when there are little or no rewards for doing so and by doing so they may be taking time, energy and focus away from the research that led most of them to their employment in universities in the first place and that will likely lead to advancement of both their field and their own careers.

**Using SETs for public scrutiny of teaching/subjects**

When SETs are to be used in this way, they should be summative and the focus should be on overall teaching competence. The fact that the audience for these will be public begs the question of how accurate or useful these SET results are likely to be. Which universities are likely to make public significantly poor results about a teacher or subject? If such a prospect is initially attractive to particular types of managers who like to use the ‘carrot-and-stick’ approach to managing staff, the attractiveness will fade when the realities of the situation are examined.

Where it has a say, a university is unlikely to allow negative SET results to threaten its reputation by making information about teaching and related weaknesses within an institution public. If universities are forced to make summative SET results public, they are likely either to be selective, to reflect the teaching excellence that exists within every institution or sanitised to the point of meaninglessness to hide or dilute the deficiencies and problems that also exist in every institution.

Whatever the intended purpose of the SET outcomes, a number of aspects of the ways in which they are collected are crucial to their usefulness. These are discussed below.

**Validity and reliability of SETs**

Harvey (2003) argues that students are “…important stakeholders in the quality monitoring and assessment processes and it is important to obtain their views” (p. 19). While it is difficult to argue against this claim, one does not have to probe very far into the opinion of teachers in the Australian higher education system to uncover disquiet about using SET instruments in universities to gather student views. Many academic staff with teaching responsibilities view SET instruments with suspicion and their use by management with scepticism and perhaps, cynicism. However, according to Feldman (1997), conclusions from twelve research reviews conducted between 1977 and 1992 indicate that the following generalisations are unsupported by evidence:

- Because of their immaturity and lack of experience, students are unable to make consistent judgements about teaching;
- Students are unable to make accurate judgements about teaching and teachers until they are away from the university for several years;
- Student ratings schemes are essentially popularity contests with the friendly, humorous teacher often the winner;
- Student ratings are both unreliable and invalid.

Despite the lack of empirical support for these claims, anecdotal evidence indicates that many academic staff believe one or more of these statements to be true. However, in some cases, particularly where insufficient attention has been paid to the development of the instruments, caution may be warranted. As Harvey (2003) comments, student evaluation of teachers “…tends to be a blunt instrument…” (p. 17) and, he says, while it may be able to identify very poor teaching but may not be very useful for the continuous improvement of teaching. However, Harvey’s (2003) comments may only be accurate for a particular kind of SET, as discussed further below.

**Development of SET instruments**

Unfortunately, in Australian universities, student evaluation of teaching instruments are often, and perhaps typically, developed without appropriate attention to the psychometric properties of the instrument. It is not uncommon for a number of staff in a university to contribute suggested items or questions to a bank, from which some or all may be drawn to make up an instrument. The items or questions may be related to the teacher, the subject/course, the environment, facilities, resources, the provision of ICTs and any other factors in any combination. Often, the measure of an element of the student’s experience is from a single item or question, rather than a scale containing a number of items or questions. Items, questions and whole instruments are rarely piloted and normative data almost never compiled. Harvey (2003) suggests that because of the typical method through which SET instruments are developed, the outcome is often “…a bland compromise, designed by managers or a committee, that serves nobody’s purposes” (p.17). Of particular concern are the reliability and validity of the instruments.

**Evidence of the reliability and validity of SET instruments**

In relation to student evaluation of teaching, McKeachie (1990) concludes that the research evidence indicates that students are generally good judges of teaching. He goes on to say that this is surprising given that most of this research has been carried out at first year level where it might be expected that students would be less able to evaluate teachers than later year students.

Based on his review of a number of relevant studies, Paulsen (2002) concludes that the reliability of student ratings of teaching is generally robust. He reports that inter-rater reliability coefficients vary according to sample but are around .70 or higher on typical questionnaires such as the Student Evaluation of Educational Quality (SEEQ) (Marsh, 1987). Paulsen (2002) adds that reliability coefficients for stability over time of student ratings are also strong, with ratings at an average of .83 (Marsh and Dunkin, 1997). Paulsen (2002) cautions that for summative purposes in particular, student ratings for individual teachers should be collected from sufficiently large samples.

Marsh and Roche (1994) point out that SET instruments, as a measure of the hypothetical construct of ‘teaching effectiveness’, are difficult to validate since a single criterion of effective teaching is insufficient. They advocate a construct validation approach. To establish construct validation, one’s construction of a particular issue must agree with other constructions of the same underlying issue (Cohen, Manion and Morrison, 2000). Marsh and Roche (1994) report that SET ratings relate to the ratings of former students and teacher self-evaluations of their teaching both significantly and consistently. Based on his review, Paulsen (2002), reports that that research findings related to the relationships between student ratings and other indications of teaching effectiveness such as self and peer ratings and qualitative student evaluations provide support for the validity of student ratings in the evaluation of teaching.

According to an extensive thorough review of the SET literature reported in Marsh and Roche (1994), appropriately measured class-average SETs are multidimensional; reliable and stable; primarily a function of the teacher of a subject rather than the subject that is taught; relatively valid against a variety of indicators of effective teaching and relatively unaffected by a variety of variables hypothesised as potential biases.

But the instruments to which these researchers refer are well-developed, psychometrically sound questionnaires, refined over many years by experts in the development and refinement of instruments. The SEEQ (Marsh, 1987), and its Australian version, the ASEEQ (Marsh and Roche, 1994) have established reliability and validity and therefore can be guaranteed to gather meaningful data on teaching consistently over time. The problem is that ‘home made’ SETs are being used in many Australian universities without the necessary psychometric bases and assurances and their results are often treated and used as if they are valid and reliable.

As one example, one university in which the author of the current paper has worked in the past, teaching effectiveness was monitored by the university teaching and learning committee charged with ensuring quality assurance. While further data and information were available, the committee essentially focused on student responses to one item in the compulsory teaching evaluation questionnaire. The item read, “This subject was well taught” and students indicate the strength of their agreement or disagreement with the statement on a five-point scale (Strongly Disagree (1), Disagree (2), Neither Agree nor Disagree (3), Agree (4), Strongly Agree (5)). Results were then calculated on the mean score for each item. Despite the fact that teaching and learning are multi-dimensional and the quality of both potentially affected by a very large number of influences and factors, including multiple teachers in many courses, the individual teacher in this university must ‘wear’ this single number measure of their teaching. This is not to suggest that this university’s approach should be criticized in particular – the type of procedures described here occurs in a number of Australian universities.

As Marsh and Roche (1994) point out, SET surveys “…should contain distinct groups of related items…derived from a logical analysis of the content of effective teaching and the purposes that the ratings are to serve, and that are supported by theory, previous research, and empirical procedures such as factor analysis and multitrait-multimethod analysis” (p. v). Often, SET surveys used in Australian universities do not meet these basic criteria. And even if the reliability and validity of SETs can be guaranteed, as Theall and Franklin (2000) suggest, given the pace of change in postsecondary education, continual revision of our approaches is necessary to ensure that student ratings data are valid and reliable in our ever-changing contexts.

Interpreting SET outcomes

Handing out SET data, even if undertaken with appropriate attention to the aspects outlined in Table 1 and discussed above, is insufficient without offering the SET audience support in interpreting the data. A number of Australian universities offer guides and tips on how to interpret SET data but as Theall and Franklin (2000) point out, there has been little systematic study of what works what does not work and that the evaluation of even anecdotal evidence would be a good start.

Harvey (2003) proposes that in order for views collected from students to be effective in quality improvement, data collected from questionnaires and the like must be transformed so that it can be easily used within a university to bring about change. He suggests specifically that student views should be integrated into “…a regular and continuous cycle of analysis, reporting, action and feedback” (p. 4). He proposes that action should not only take place on the basis of SET outcomes, but be seen to have taken place. This suggests the delegation of responsibility for action; accountability for action and inaction; the commitment of appropriate resources (Harvey, 2003) as well as careful attention to the issues around the communication of SET outcomes.

Conclusion

Harvey (2003) suggests that the agenda for changes to teaching based on SETs must be progressive and not recriminatory and that action should improve the student learning experience. He adds that achieving all of this is difficult and may help explain why, despite the best intentions, data gathered from students is not always used to bring about positive change to teaching and learning.

Should the DEST proposed changes around the use of SETs be adopted in Australia, there are a number of ways in which universities might respond. One is to adopt a one-size-fits-all model, designed by a well-meaning university committee, that focuses on the communication of a positive spin.
on teaching to the public and the communication of disapproval to staff who fail to meet standards or benchmarks, however these might be determined. Issues of reliability and validity will not be paramount with such an approach. An alternative is for universities to genuinely and systematically address teaching quality, the results of which will subsequently be reflected in the publicly communicated information. Such teaching quality could be achieved through the use of valid, reliable SETs; through the provision high quality diagnostic and developmental programs and processes for staff that are based on the evidence from these SETs and through the development and implementation of probation, increment and promotion systems that genuinely reward excellent teaching.

References


