International Journal for Educational Integrity

Student Evaluation of Teaching

Peter Slade
University of the Sunshine Coast
slayed@bigpond.com

Chris McConville
University of the Sunshine Coast
cmcconvi@usc.edu.au

Keywords: student evaluation of teaching, validity, higher education

Abstract

This article considers the validity and usefulness of student evaluations of teaching (SET) at a small Australian university. Face and content validity were considered and a factor analysis was performed to evaluate the overall validity of a survey instrument which purports to give useable results in respect to teaching methods and approaches. It was found that the survey instrument was flawed in that the ten compulsory questions of which it is constituted, all collapsed into one dimension. This dimension was determined to be the extent of popularity of the lecturer for whom the survey was conducted. In essence, the survey is not an evaluation of teaching, but rather students' opinions of the lecturer concerned. It was concluded that the SET survey serves no educational purpose and is a violation of academic freedom and lecturers' rights.

Introduction

Universities throughout the western world now routinely impose on both academic staff and students, some evaluation of teaching standards through surveys. In Australia, reviews of higher education under the Howard Liberal government have resulted in a compulsory national system through which students assess their lecturers. Results from these student evaluations are now incorporated into rankings of teaching quality for Australian universities and available through the Commonwealth government's website. These ostensibly provide prospective students with a comparative grading by which they can make judgements about teaching quality in different degree programs at various universities.

These systematic surveys depart from the long-standing practice of lecturers who informally gathered responses to course content and style. University teachers habitually surveyed their students, usually in an attempt to improve their teaching effectiveness. For example, they may have wished to gauge responses to a particular mode of instruction or segment of course content. On the other hand, some teachers posed more generic questions (and questionnaires) about what students thought of particular courses. In addition, staff have often used some form of class observation by mentors, peer review, analysis of student drop-out rates and other indicators in order to improve classroom strategies. However, in the current Australian university system, SET holds a central place in measuring the teaching standards of individual lecturers.

No doubt in their informal nature, individualistic surveys and peer review have failed to uniformly address issues of teaching quality and student learning, focussed as they were on the individual concerns of the lecturer. Hence the appeal of SET. Lecturers

The International Journal for Educational Integrity is available online at:
http://www.ojs.unisa.edu.au/journals/index.php/IJEI/

who devised their own surveys were interested in educational methods and outcomes intrinsic to their discipline. The statistician's need to ensure validity and reliability had little impact because the sole reason for conducting surveys was to improve a particular course. Comparisons across lecturers, faculties, disciplines and universities were of little importance, and the surveying of students for their evaluations was not mandatory across the higher education sector as a whole.

It is easy to see why the current SET survey system appeals to Commonwealth ministers of education. Between 1950 and 1995, when governments underwrote a massive extension of universities throughout Australia, they did so with some notion of nation-building, civic responsibility and the intrinsically progressive social consequences of education. More recently, such imperatives have given way to a market-oriented understanding of education. According to this view, Australian higher education is integrated into the consumer services sector and the individual consumer's (formerly student's) pursuit of wealth through training, drives the character of her or his university education. Since these consumers now pay for their education they, as in their purchases of health insurance or package holidays, need to feel satisfied both at the point of sale and once the service has been consumed. SETs thus are of critical importance to governments in transforming the nature of higher education. If the criteria for education quality can be reduced to the satisfaction levels experienced by the consumer (student), it becomes much easier to persuade the public, (the broad mass of potential consumers) that they are buying a good product. If there are failings anywhere in the system these can be presented as the fault of individual service providors (lecturers) and not of the system itself, nor of the government, which still partially funds universities. Based on such an understanding, the intellectual content of courses, and the skills and values which a student acquires at university become secondary. The role of the student in education is also diminished and like the holiday-maker who finds foreign travels unenjoyable, can allocate all fault to the service provider and not the consumer.

SET thus appears critical in the transformation of higher education over recent years. If it is to be trusted by potential consumers as a guide, however, it needs to be reliable and valid. The recent imposition of compulsory, generic and standardised Student Evaluations of Teaching (SET) across the whole sector brings questions surrounding validity and reliability into sharp focus. There are several reasons for this.

Firstly, standard SET surveys raise questions about the ability of such instruments to elicit responses about teaching effectiveness across all teachers and disciplines. Do they allow comparisons between teachers and disciplines which are valid and reliable? Indeed, some writers, such as Becker (2000), have grave reservations about the use of survey statistics for inter-instructor comparisons.

Secondly, given the use of a national website showing SET-type results, the prospective user of such information would need to be assured that it was accurate, both in and of itself, and as a means for making comparisons. No university would want to attract (or deter) students through inaccurate assessments of teaching quality.

Thirdly, and associated with the second point above, questions arise regarding the confidentiality and anonymity of SET data. Within individual universities, open-ended questions eliciting free-response answers from students do regularly result in pithy and sometimes insightful comments. At the same time student written responses raise concerns about potential defamatory comment. These become far more serious if such commentary is widely distributed. It would be unfortunate indeed if any university exposed itself to some future legal action, through having placed contentious student comment about its employees on the world-wide web.

Fourthly, there is the question of possible uses of SET beyond the improvement of teaching standards. SET can become a convenient and not necessarily reliable guide for promotions committees. In extreme cases the results might offer an easy tool for justifying retrenchments on the one hand, and fast-tracked promotion on the other. Any organisation would want to be assured of the SET instrument's reliability and validity before relying on such surveys for staff management. For, as a beguilingly straightforward instrument for ranking teaching staff, SET's nature and indeed shortcomings, may one day become central to appeals, by both failed promotion applicants and retrenched staff.

Fifthly, and most importantly, lecturers would like to make use of SET results as an aid to improving or changing their teaching. Becker (2000) raised this point when writing about the improvement of teaching in Economics. Obviously, there is little point in having a SET instrument that offers no specific suggestions for teaching improvement. The mandated and generic instruments ought to be an aid to lecturers in capturing specific or fine-grained information about their particular teaching style, methods and discipline. In essence, this last point leads us to issues of academic freedom and teaching quality. Mandatory SET instruments could have a bearing on the way in which teachers go about the scholarly business associated with a discipline. SET's influence might not prove positive. In language teaching for example, repetition and rote learning may well have a valid role, but one which can bore and frustrate students. If teachers of languages are persuaded by low SET scores, to reduce this activity, SET could be seen as having a negative impact on learning quality; a direct result of its generic and non-specific character.

SETs take up valuable resources and time in universities. If they are to have real value as guides for prospective students, for staff keen to improve teaching and for university administrators deciding which of their academic staff to sack, mentor or promote, SET needs to be both valid and reliable. The first question we need to ask is, "Just how valid and reliable are the SETs currently in use in our universities?" Once this matter has been addressed, it ought to be possible to extend or alter survey instruments so they can respond sensibly to the points raised above. In this article, we consider the validity, and to a lesser extent, the reliability of SET. We use as an example an instrument currently in use in one regional university in the Commonwealth of Australia. For reasons of anonymity, we call the regional university, "Fisher University".

Process Evaluation

Any attempt to measure and evaluate a process, whether of production, a social system (which many production systems are), of nature, or indeed in the physical sciences, demands that the process be understood in the first instance. For example, in a physical production system, all inputs, processes and outputs have to be understood, and it should be possible to measure and therefore evaluate them. Tolerances within a range of possibilities are allowed for. It is relatively easy to understand what the various facets of a physical production system are, and it is relatively easy to measure them. Having prescribed a set of standards for the quality of the facets, evaluation from the measurements is then straightforward. If some aspects lie outside limits of tolerance, the process can be adjusted so that the faulty items are no longer produced. Thus, objective data about the process are collected (measurements) and analysed to ensure an acceptable standard of quality. J. Edwards Demming (1986) once stated, "In God we trust, all others require data."

If teaching is to be regarded as a type of production process, as the use of SET-like instruments implies, then it is necessary to understand the system of teaching in its entirety; all inputs, the process and outputs have to be considered and then a range of desirable dimensions can be promulgated. These need to be objective, clear and able to be measured. Given any drift in the system, it should be possible to take corrective action to rectify any unsatisfactory aspects of the process.

In point of fact, setting up a system of evaluation of teaching requires a very deep consideration of selecting and validating measurements and their dimensions.

Broadly, the evaluation requires the following steps. Firstly, an operational definition of the construct(s) or phenomenon(a) would be necessary. For instance, what exactly is learning or teaching, and is it possible to know when it has occurred? Next, those desired characteristics within the phenomenon(a) need to be identified. This might be “independence of thought” or “advanced reasoning ability”. The last major step is to identify and examine relationships between possible measurements, the characteristics associated with the phenomena and the characteristics of other (perhaps extraneous) related phenomena. This last step ensures that processes closely involved with teaching but which are unimportant or undesirable are not accidentally measured. These confounding variables have to be clearly understood and removed.

In summary, it is necessary to know what is being talked about and how measurement might be carried out. Overall then, the purpose is to develop a theory about the relationship between the phenomenon(a) to be measured and the measurements themselves. An important consequence of such safeguards is that characteristics of high quality teaching would not be automatically associated with low student ratings in a SET survey.

Reliability and Validity

Whenever an instrument is designed, two fundamental questions have to be addressed to ensure that it is useful. The first concerns the extent to which the instrument returns similar results through time and in different situations. The second is concerned with the extent to which the instrument reflects information on the concept, variable, construct, etc. That is being considered. These two aspects are not necessarily separate and distinct from one another, but are rather, associated with each other.

The first concern is known as reliability. Reliability has been suggested as being, “the extent to which a variable or set of variables is consistent in what it is intended to measure. If multiple measurements are taken, reliable measures will all be very consistent in their values” (Hair, Anderson, Tatham, & Black, 1984). Thus, through time, in varying circumstances and with different subjects, a measure will not drift and will remain consistent. The other concern, validity, has been said to be, “The extent to which a measure or set of measures correctly represents the concept of study (or idea under consideration) - the degree to which it is free from any systematic or non random error” (Hair, Anderson, Tatham, & Black, 1984). In other words, this is concerned with whether or not the instrument measures what it is supposed to measure. Hair, Anderson, Tatham and Black (1984) compared validity and reliability by stating, “Validity is concerned with how well the concept is defined by the measure (s), whereas reliability relates to the consistency of the measure(s)”.
In general it is not possible for an instrument to be unreliable and valid, although in some cases it is possible for an instrument to be reliable but invalid. In the last case, it is possible to collect consistent data (through time and space) about an entirely incorrect, misunderstood or inappropriate construct or variable.

In light of these definitions, it could be proposed that a valid SET-type instrument might never exist. Martin (1998) has argued along these lines, giving a number of reasons for his view. It is not possible to find an explicit operational definition of high-quality teaching. This is because all current processes which rely on student evaluations to gauge teachers or teaching, in fact use students’ opinions to define high-quality teaching. Moreover, even if an operational definition were available, students are unlikely to be the most suitable candidates to judge many of the aspects involved (Caskin, 1983; Reckers, 1996). It is illogical to expect that a typical student could judge the currency and relevance of knowledge gained in a particular subject area. Is there understanding on the student’s part about any theory that may underpin a set of knowledge (Martin, 1998, p. 1080). Lastly, there remains the issue of the incomparability of performance evaluations. Demming (1993) always maintained that

94 percent of the variation in any system could be attributed to the system itself, and not to the people working in the system. Therefore, different results from different individuals carrying out the SET survey would, in all probability, result from factors associated with the system and have nothing to do with these individuals.

Furthermore, system variation is not equally distributed across workers (in this case lecturers). The counter assumption (on which SET relies) stems from the mathematical notion of the existence of a normal distribution and that variation within that distribution is randomly spread. The assumption is not borne out by reality. Within any system, some components are prone to have a greater variation than others. For example, some parts of machines are more prone to drift and breakdown than others. So it is with any system of production, and in teaching and learning. There is no logical reason to suppose a random distribution of variation in teaching, nor is there any empirical evidence to suggest there is. Demming (1993) wrote, “Ranking a group of workers is merely an exercise in ranking the effects of the system on the workers.”

To apply this dictum to our universities: in educational situations, many crucial constraints are beyond the lecturer’s control. These may include inadequate resources, poor equipment, funding cuts, lowered entry standards, poor teaching time scheduling, and so on. Such resources may be unequally distributed across subjects and faculties. All of these contribute to the variation within the system. They have almost nothing to do with the lecturer.

On the other hand, irrelevant or extraneous aspects of lecturers can influence the results of SETs. Feldman (1986) reported that the overall relationship of instructor personality with students’ ratings is substantial (Feldman, 1986, p. 38). Felton, Mitchell and Stinson (2004) cite a website that asks students to rate the “sexiness” of lecturers. They argued that such a dimension is irrelevant in assessing the worth of a lecturer, yet they found that about half the variation in rated lecturer quality is a function of “easiness” and “sexiness” (Felton, Mitchell & Stinson 2004, p. 92).

Many studies have considered the impacts of extraneous or confounding variables on the results obtained from SETs. This points directly to the issue of validity. Davell and Neal (1982) have found that validity coefficients in all the studies they examined are so variable that a meaningful and generalisable estimate of their validity does not exist. Abrami, d’Apollonia and Cohen (1990) concluded that whereas the average validity coefficient for global ratings is moderately positive, the results of their studies appeared inconsistent both from study to study and across rating dimensions. Cashin (1990) found that high student ratings occurred in the arts and humanities, although English literature and history fell into the medium to low category. Low rating occurred in business, economics, computer science, mathematics, physical sciences and engineering.

Such variation might explain why there is a commonly held belief that SET surveys are not all that different from popularity contests. This belief seems to be borne out by research. Arreola (1995), Aleamoni (1987), Feldman and Neal (1990) and Franklin (1990) all reached this conclusion having carried out meta analyses and having reviewed hundreds of studies dating back to 1921. Furthermore, it has been found that generally there is little or no correlation between student achievement and student ratings. Cohen (1983), McCallum (1984) and Damron (1996), agreed on the likelihood that most of the factors contributing to student ratings of teachers and teaching are unrelated to the teacher’s ability to promote student learning.

Yet SET surveys are now compulsory across Australian universities and many university administrators seem oblivious to their methodological problems. Irrationalities inherent in translating statistical analyses of production-type processes to learning and teaching seem to have barely registered. Very often SETs are conducted with little or no consideration for statistical problems, data adequacy and proper interpretation of data. For example, small data sets, low response rates, high response rates, and the use of global ratings are often not properly worked through. (Emery, Kramer & Tian, 2003; Martin, 1998). The most obvious fault lies in the representativeness and adequacy of students answering surveys (Isley & Singh, 2005). Students who may have attended very few classes are given an opportunity to answer questions about the entirety of a course. Students are asked to make judgements about how much they have learnt from a course of study before their final results are available. Their answers are thus based on a guess, often one that proves wildly inaccurate.

It is reasonable to assert that at a more fundamental level, students are not properly qualified to evaluate teachers and teaching. Adams (1997) raised the important point that students, "…universally considered as lacking critical thinking skills, often by the administrators who rely on students evaluations of faculty are able to critically evaluate their instructors." Clearly this is illogical. In nearly every situation in which people are expected to evaluate systems, production processes, machinery, and even art, they are trained to make judgements. Yet in the case of SETs students are given carte blanche to make judgments upon the professional abilities of those, who by the very fact of having been hired to do the teaching, are recognised by universities as better understanding course content and teaching strategies than their students.

There is ample evidence to suggest that the construction and use of SET-type instruments, along with their administration, is flawed. For example, when faced with the necessity to introduce a system of SET, one provincial Australian university adopted the entire system used at a nearby metropolitan institution; one with a far different staff and student profile. This example of SET, initially proposed as voluntary and provided to individual staff members in a confidential manner, was later adopted as compulsory and available to line managers. As part of the introduction of this system, it was stated that lecturers owned the results of the surveys and need only make them available to third parties at their discretion. For example, it was stated that should any lecturer wish to apply for promotion they would be required to present all their SET results to the promotion panel. After a year of using the system, results were then deemed available to the Deputy Vice Chancellor and Deans, in addition to the lecturers themselves. There are examples of staff proposing alternative mechanisms for student evaluation. These have been rejected by a teaching committee. Staff are thus stuck with the one system which is becoming used for career-related decisions. For whatever reason the university union body agreed to these changes.

There is, unfortunately, a real danger that a university adopting such processes might expose itself to legal challenge. It has been shown that, in the USA, parties performing ratings who are not qualified or do not possess a strong common interest, and where privilege does not exist, are open to defamation suits (Colson V. Stieg, 433 N.E., 2nd, 246, [III 1982]). Whilst defamation laws in Australia are complex and vary from state to state, students often tread a very fine line in response to the open-ended questions in SET. Where uncensored results are placed on a website or used in promotion decisions by universities, issues of unfair and derogatory judgements about professional ability may well bring us to the brink of slanderous comment. An example of a SET instrument used in one university is the basis of this paper. By exploring the statistics produced by SET rather than debating the legalities of its compulsory imposition or the possibility of compensatory legal action arising out of SET, this paper provides a basis for reviewing the central educational role of SET in a broader fashion and with an eye to better evaluating teaching standards.

The SET system discussed here was comprised of ten compulsory statements about which students were required to make comment, by entering their opinion on Likert five point scales. In addition, lecturers could select a further ten optional statements from a database of around 100 questions. They were entitled to insert one statement of their own construction.

Under current arrangements lecturers with 0.5 appointments or above are required to use SET in one of their courses each year. The survey forms are passed onto the university information technology section and ultimately processed at another university. The results are returned to lecturers in a standardised form, showing histograms and mean scores for each question. Some lecturers, hoping to perform a more fine-grained analysis of the results than that permitted by the descriptive print-outs, asked to have their results forwarded to them in electronic form. This request was denied.

Validity and Reliability in SET

The SET instrument comprises two sections. The first section, consists of ten statements to which students record their reactions on a five point Likert scale. This is compulsory and the lecturer has no input to its composition or administration. The second section, once again comprising ten statements to which students record their reactions on a five point Likert scale, is optional. Lecturers can select their statements for inclusion from a database of one hundred questions.

There are several points that should be made about the statements from both sections. The ten compulsory statements (as do most of the other questions) lack face validity. The instrument purports to be a survey of student feedback on teaching, yet all the statements deal with opinions about the teacher. Thus, even at a very fundamental level, the survey instrument is flawed. There are further serious problems with each of the ten statements, and these are dealt with individually below.

Statement 1: The lecturer makes clear what I need to do to be successful in this unit.

It is usually a presupposition that students undertaking a course of study are cognisant with the idea that there is a corpus of work with which they must be familiar Therefore, this is a difficult statement to comprehend, because its meaning is unclear and it can be interpreted in a number of ways. Consider the following interpretations.

Most university teachers have run into the annoying problem whereby students ask, “Is this going to be in the exam?” The question is best not answered. However, SET Statement One serves to encourage students in the belief that such a query is a legitimate part of higher education. Clearly, in a conventional sense, students are made aware of the course assessment methods and requirements; what each piece of assessment is, what it represents in total course marks, and further requirements such as students should achieve fifty percent or better to gain a pass. In many courses, students were supplied with a detailed marking scheme for each item of assessment. Under these circumstances it is difficult to know how students could have done other than to have marked SA (strongly agree), or a value of five in the Likert scale.

It is possible that intervening and confounding variables could enter into the process. For example, students who scored badly in the course work, and are unable to take responsibility for their poor results, might blame environmental factors as a rationalisation for theses outcomes. Alternatively it might simply be a dislike of the teacher concerned. And this is one of the crucial points; the survey is fundamentally a survey of opinion, not some verifiable facts.

Statement 2: The lecturer is skilled at developing a class atmosphere conducive to learning.

This is also a difficult statement to comprehend. The questionnaire does not stipulate whether “atmosphere” refers to physical, social, emotional or other environment. Without a knowledge of what constitutes a “class atmosphere conducive to learning”; it is not possible for people to evaluate a teacher’s performance in this regard.

Moreover, such a class atmosphere would probably be highly peculiar or idiosyncratic to each student concerned. Some people might prefer a highly structured environment, whilst others might prefer a more relaxed arrangement. Even so, how people might think they learn best is necessarily the circumstance in which they do learn best.

In so-called learner-centred environments, it is the students who set the tone, and these circumstances may not suit all within those circumstances. It is impossible to make any use of opinions garnered about “atmosphere”.

Statement 3: The lecturer has a good manner (eg friendly, helpful, and enthusiastic).

The usual problems are evident; students are not qualified to evaluate the statement, and it could mean different things to different people. A “good” manner is not likely to have a positive impact upon learning. It is not necessarily the case that friendliness inevitably leads to important educational outcomes. Perhaps a challenging manner, which does not necessarily accept substandard effort and sloppy thinking, is more likely to ensure both learning and poor student evaluation.

To take an example from outside tertiary education: Military recruits are treated in an almost hectoring and bludgeoning manner as they learn the rudiments of military life. In fact in the Australian military, there have been cases of torture and bullying, if recent accounts are to be believed. Such excesses apart, the imposition of military discipline appears, in the main, to train people extremely well for the rigours of battle and for the uncomfortable circumstances of an outdoor and physically demanding life. It would be instructive to know how recruits rated their drill and physical training instructors. Similarly in elite sports organizations, coaches (teachers) subject their charges to rigorous and highly competitive training. The nastiest of coaching staff can have the best (as rated objectively in win-loss ratios) learning outcomes.

Statement 4: The lecturer shows appropriate concern for student progress and needs.

Confusingly, this statement does not define “appropriate” and any meaning would vary from person to person. In higher education, student progress is in the hands of the students themselves. In fact an appropriate concern might well be one which is harsh and unsympathetic since such an attitude could easily produce greater student commitment, even if only to spite the nasty lecturer. The statement is similar to others in the questionnaire in as much that a lecturer scoring poorly in say, statements two and three, ought not score badly in four. This is because a lecturer, showing concern for students’ needs and progress, might institute an atmosphere of deliberate austerity and might demand great efforts from students, so as to ensure an improvement in their academic progress. Bertrand Russell once said that there is no high road to algebra (Russell, 1976). Yet there is evidence that this is contradicted in the surveys. Thus, despite a high level of redundancy built into the questionnaire, it is apparent, given a lack of training, that students could still give unreliable answers.

Statement 5: The lecturer provides feedback that is constructive and helpful.

Once again, the issue of redundancy arises. It would seem to be impossible to suggest that the definition of a class atmosphere that is conducive to learning would not have feedback as one of its elements. Yet the questionnaire presumes the two are not related. The same applies to the concern for student progress.

The potential for confounding variables is present as well. Sometimes cogent feedback requires an honest statement of shortcomings, as well as indications as to where improvement might be made. Some students would feel offended by such honesty and may tend to rate a lecturer poorly.

More importantly, there is mounting evidence that increasing the amounts, detail and timing of feedback is not necessarily “beneficial” to students. Studies have demonstrated that students provided with less feedback, or feedback given in summary form, perform less well in immediate situations, but perform much better in the long run (Hunt, 1997). Thus, asking students to judge the usefulness of feedback, in respect to that offered by a teacher is, at best, invalid.

Statement 6: The lecturer helps me to improve my understanding of concepts and principles.

Redundancy is, once again, a problem. For example, it would not be possible to have a good learning atmosphere whilst simultaneously not having an improvement in understanding concepts. Moreover, it is not possible for students to know whether or not they had understood concepts directly as a result of the efforts of the lecturer. Intervening and confounding variables might include whether or not the lecturer was an extrovert and the presence of “gaming”. Becker (2000) has raised issues surrounding the temptation to, and the means by which, lecturers might improve their SET scores by tactics, which are not related to teaching. These could include giving students test answers, and getting rid of troublesome students prior to the administration of a SET instrument. Obviously, such variables are not controlled for. Generally then, concept formation and understanding are usually experiences gained in individual circumstances, given self-discipline and work, yet the statement presumes the lecturer is the primary agent in the student’s understanding. Finally, what do we do about the increasing number of subjects in which conceptual understanding is no longer central? As universities become more vocational in their programs and students more concerned with so-called “practical” knowledge, the acquisition of concepts and underlying principles can seem arcane and irrelevant.

Statement 7: The lecturer structures and presents the unit content in ways that help me to understand.

The two “perennials” of redundancy and the inability to judge are present. It would not be possible to score highly on this statement and poorly on statement 6. Further, it is not possible for students to judge whether or not a particular presentational form helped them to understand. This is mainly because of the presence of confounding variables, once again. It is possible that less capable or lazy students might not understand content, but that is not something that should be “blamed” on the lecturer.

There is also the problem of the negative being irrational. No lecturer would deliberately present unit content in ways which are difficult to understand. The job is about communication, and teachers generally strive to improve in this regard.

Statement 8: The lecturer is knowledgeable in their subject area.

This statement is simply unanswerable. If students could rate the lecturer’s knowledge against the global knowledge in a discipline then they would be foolish to waste money by enrolling in the course. If they were so foolish they would make very unreliable respondents. It is somewhat fanciful for universities to expect students to make this type of judgement. It completely ignores the Dr. Fox effect and the presence of all sorts of confounding variables.

Statement 9: The lecturer sets tasks that are useful as learning experiences.

Tasks set may or may not be useful as learning experiences, but students may not be sufficiently capable of learning from them. Moreover, given the implied incapacity of providing honest feedback in the statement, the students would not be able to tell whether or not they were learning anything.

Statement 10: Overall, how would you rate the teaching of this lecturer in this unit?

Such global statements are inherently invalid. There are far too many intervening variables. Furthermore the notion of rating is comparative. Against whom is the lecturer being rated? What are the criteria (emotional affinity apart) on which the rating is being made?

Factor Analysis

Many Australian university staff share concerns about the reliability and validity of SET instruments, of which this is but one example. To address some of these concerns a factor analysis was carried out in respect of students' responses to the ten compulsory questions.

Broadly, factor analysis, or more particularly in this case, principal components analysis, enables the definition of an underlying or latent structure in a data matrix or data set. It facilitates the analysis of the structure of the interrelationships (correlations) among a large number of variables by defining a set of common underlying dimensions, usually called factors. Thus, it is possible to reorientate the data so that the first few dimensions account for as much of the available information as possible. If there is much (or any) redundancy in the data set, then it is possible to account for the most of the information in the original data with a considerably reduced number of dimensions.

Teaching staff were asked to contribute their students' responses to the authors, who had these punched into machine-readable form, remembering that Fisher University itself did not supply electronic responses to lecturers. In all cases anonymity was guaranteed, and all response sheets were returned to the staff concerned. In all, enough teaching staff contributed their responses such that a sample size of 625 student responses was available. The subject areas in which staff taught covered a wide cross-section, including law, economics, history, finance, and marketing.

The survey statements were used as the input variables into the analysis, and were coded according to the following schema.

Table 1: Variable Meanings and Codes

Statement number and statement	Variable Name
1. The lecturer makes clear what I need to do to be successful in this unit.	Success
2. The lecturer is skilled at developing a class atmosphere conducive to learning.	Atmosphere
3. The lecturer has a good manner (e.g. friendly, helpful, and enthusiastic).	Manner
4. The lecturer shows appropriate concern for student progress and needs.	Progress
5. The lecturer provides feedback that is constructive and helpful.	Feedback
6. The lecturer helps me to improve my understanding of concepts and principles.	Improve
7. The lecturer structures and presents the unit content in ways that help me to understand.	Understand
8. The lecturer is knowledgeable in their subject area.	Know
9. The lecturer sets tasks that are useful as learning experiences.	Learn
10. Overall, how would you rate the teaching of this lecturer in this unit?	Overall

The following results were obtained from a principal components analysis.

Examination of the correlation matrix (Table 2) reveals that the lowest correlation coefficient was .346, being that for “manner” and “success”. The highest was .676, being that between “overall” and “atmosphere”. It should be noted that apparently no correlations stand out as either being extremely high or extremely low. Indeed, the correlations seem to cluster around a mean of about .500. This would lead to the prediction that, in all probability, only one underlying latent root typifies the data set.

Table 2: Correlation Matrix

	S	**A**	**M**	**P**	**F**	**I**	**U**	**K**	**L**	**O**
Success	1.000	.575	.404	.522	.582	.589	.589	.478	.504	.622
Atmos	.575	1.000	.591	.526	.563	.575	.607	.452	.526	.676
Manner	.404	.591	1.000	.552	.496	.446	.419	.460	.346	.563
Progress	.522	.526	.552	1.000	.625	.544	.477	.384	.442	.554
Feedback	.582	.563	.496	.625	1.000	.592	.537	.445	.493	.600
Improve	.589	.575	.446	.544	.592	1.000	.644	.461	.554	.608
Underst	.589	.607	.419	.477	.537	.644	1.000	.538	.611	.668
Know	.478	.452	.460	.384	.445	.461	.538	1.000	.471	.560
Learn	.504	.526	.346	.442	.493	.554	.611	.471	1.000	.595
Overall	.622	.676	.563	.554	.600	.608	.668	.560	.595	1.000

Examination of Table 3 showing the KMO test for sampling adequacy, as well as Bartlett’s test of sphericity, leads to the conclusion that there is a high degree of inter-correlations among the variables, and that therefore, the factor analysis is very appropriate. Broadly, KMO has a range from 0 to 1, with 1 being excellent and 0 indicating no correlations. Thus .938 can be described as “meritorious”. Bartlett’s test, which tests for the presence of correlations among the variables, returned a significance level of p <.0001. This means that the null hypothesis that there are no correlations among the variables, can be rejected.

Table 3: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.938
Bartlett's Test of Sphericity	Approx. Chi-Square	3352.062
	df	45
	Sig.	.000

Communality shows the total amount of variance an original variable shares with all other variables included in the factor analysis. Table 4 indicates a range of communalities from .463 to .723. Generally most of them lie in the .6 region. This leads to the conclusion that most of the variables have a moderate degree correlation with each other, taken as a whole. This adds to the belief that probably only ONE latent root exists in the data set.

Table 4: Communalities

	Initial	Extraction
SUCCESS	1.000	.595
ATMOS	1.000	.643
MANNER	1.000	.468
PROGRESS	1.000	.540
FEEDBACK	1.000	.607
IMPROVE	1.000	.627
UNDERSTD	1.000	.645
KNOW	1.000	.463
LEARN	1.000	.527
OVERALL	1.000	.723

Table 5 shows component variance loadings. This table indicates that 58 percent of the total variance is explained by one component. If each variable accounted for the same amount of variance, then each would explain 10 percent of the total variance, and would have an eigenvalue of 1. In this case, one component carries the variance explanation weighting of near 6 components if the variance were evenly spread. Adopting Kaiser’s criterion, that only those components with eigenvalues greater than one should be selected as new dimensions, then only one component can be legitimately extracted from the data set.

Table 5: Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.839	58.385	58.385	5.839	58.385	58.385
2	.804	8.039	66.424			
3	.665	6.648	73.073			
4	.523	5.230	78.303			
5	.477	4.770	83.072			
6	.413	4.129	87.201			
7	.364	3.636	90.837			
8	.335	3.350	94.186			
9	.297	2.967	97.153			
10	.285	2.847	100.000			

Table 6 shows the component matrix. This indicates the extent of correlation between each variable and the extracted component. The lowest correlation coefficient is that for “manner”, which is .684. Thus it can be said that all the variables are very closely correlated with the extracted component.

Table 6: Component Matrix

	Component
	1
SUCCESS	.772
ATMOSPHERE	.802
MANNER	.684
PROGRESS	.735
FEEDBACK	.779
IMPROVE	.792
UNDERSTAND	.803
KNOW	.680
LEARN	.726
OVERALL	.850

Finally, Cronbach’s Alpha test for reliability was run, and was found to be .9209. This is regarded as an exceptional score in that it shows that all the items in the component are highly correlated with one another.

Discussion of Principal Components Results

Fundamentally, these results show that the ten questions, treated as variables, all collapse into one component. In other words, there is a high degree of redundancy in the instrument. So much so, that all ten statements say the same thing. It is reasonable to speculate as to what this might be.

Each of the survey instrument questions/statements have been analysed in detail but some recapitulation is necessary when contemplating the meaning of the derived component. All the statements ask students for their opinions of the lecturer, and not about teaching. Hence the instrument lacks face validity. This is to say that an instrument which purportedly attempts to gain feedback on teaching, actually canvasses student opinion of the lecturer or teacher. Many of the questions/ statements ask for student opinion about aspects of the teacher for which they are unqualified, or lack competence to judge. For example, the lecturer’s knowledge, the efficacy of tasks and the global rating are outside the competence of students. Many of the statements are meaningless or undefined. For instance, opinion concerning the development of a learning atmosphere, appropriate concern for student progress and needs, the provision of helpful feedback, the lecturer having a good manner, the presentation of material to aid understanding, and helping student understanding generally, all dissolve into groundless statements. Given the amorphous nature of the survey instrument, students would be forced to either not complete it, or to engage in a process of “second best” by responding with unsubstantiated opinions about a lecturer.

Generally then, examination of the statements that comprise the instrument reveals that they comprise students’ opinions about supposed aspects of the lecturer’s teaching, which of course, is an undefined notion. It is therefore quite reasonable to assert that the instrument boils down to students’ opinion of the lecturer(s). The survey could be made more transparent by asking one simple question: “What do you think of the lecturer?” When reduced to this simple form the SET would give a clear picture of academic staff popularity and some indication of student mood. It might then be a quite useful tool in aiding teaching staff. Stripped of any pretension to objective assessment of educational processes it would become an invaluable guide to the necessary ingredients of popularity with a predominantly teenaged group of Australians. In other words, it is hard to see how SET, as is evident in this example, and probably as would be the case even with another set of questions, can be understood as anything more than a popularity poll. However, it is stunningly reliable in delivering this verdict. There can be no doubt about this. The results clearly show that in order to score very highly in the instrument, lecturers should simply aim to make themselves popular with their students. And here lies the heart of the invalidity; an instrument which purports to measure the efficacy of a person's teaching, measures, with admirable accuracy, another characteristic.

Voluntary Section: Ten Statements

Not wishing to condemn SET on the basis of the compulsory section, it may be that the ten voluntary questions allowed lecturers to return validity and reliability to the process. The second part of the survey instrument was voluntary (to lecturers) with ten questions selected from a data base. Choosing these questions was often a difficult task given their often vague and tautological character. As an overarching test of students' capacity to address statements logically and recognize irrelevancies, two of the lecturers who contributed their responses to the authors asked intentionally irrelevant or unanswerable questions as a guide to the reliability of the overall survey. Discerning students would respond with an N/A to such questions. In one case, for every statement, with one exception, at least 94 percent of respondents did not enter N/A. Indeed, the lecturer concerned continued to score quite highly in the Likert scale. Lecturers are thus being given enormous credit for effectively performing a task which in reality never happened.

Alternative Bases for Setting up SET

Imposition of this version of SET occurred at the same time as the university drew up a list of graduate attributes. As occurs in many Australian universities, lecturers are now required to incorporate these attributes into their course outlines, showing how various components of their assessment help develop attributes in students. The list of attributes is shown in Table 7.

Table 7: Graduate Attributes

Attribute	Explanation
To understand	• To have relevant, discipline-based knowledge, skills and values • To be able to apply and evaluate knowledge
To think	• To value and respect reason • To be able to reason competently
To learn	• To be self-aware, independent learners • To be able to collect, organise, analyse, evaluate and use information in a range of contexts
To interact	• To be able to interrelate and collaborate • To value and respect difference and diversity
To communicate	• To speak, listen and write competently • To be competent users of information and communication technologies
To initiate	• To be constructive and creative • To be enterprising
To value	• To have self-respect and a sense of personal agency • To have a sense of personal and social responsibility • To understand and apply ethical professional practices

Source: Fisher University 2004.

Arguments about the usefulness or legitimacy of these attributes aside, any university could develop a quite informative survey based around these and similar lists of graduate attributes. Thus, statements or questions could be put together in such a way as to find the extent to which students consider themselves to be moving in the direction of these attributes. This might prove helpful to staff, students and administrators. Given seven attributes made up of 15 “variables” it is conceivable that a series of statements testing for the variables, and assuming they would be properly constructed, would lead to the statements collapsing into 7 dimensions in a principal components analysis. Thus, a very useful instrument might be obtained. Indeed it would have a basic validity absent from the current SET. Students would now be rating their own performance rather than that of another.

Conclusions

SETs, as currently used in Australian universities, provide an interesting and informative insight into the evanescent moods of young Australians. They may be of use to universities in tracking subtle alterations in student opinion across distinct regions and over time. There remain important questions about SETs’ ability to stimulate better educational procedures. It does not appear possible to construct a valid instrument concerned with students’ evaluations of teaching. There are several reasons for this. Firstly, it is not possible to envision all the inputs, processes and outputs of teaching or how the process of education works. Even if this were possible, there would be wide disagreement as to what constituted useful, appropriate or valuable outputs or results. Secondly, in the example chosen here, the first problem was compounded by administrators acquiring a pre-existing evaluation instrument which had an inbuilt invalidity and redundancy. Thirdly, students are not trained to evaluate lecturers and the teaching they might receive, and have been seen to respond to even irrelevant statements or questions. This leads to concerns about the wider use of SET.

Given the underlying problems lying behind an apparently concrete set of histograms and tables, student evaluations of teachers may well raise awkward issues in relation to human resource management in universities. There remains the vexed matter of SET’s potential encroachment on academic freedom. Where course structures are altered in response to SET this too could be shown to be unjustifiable. A university might unknowingly respond to SET through requiring learning activities which make lecturers more popular but which hinder learning. The legal obligations of students to refrain from defamatory remarks may perhaps be avoided because of the anonymous nature of the survey. They may be less easily evaded by universities which publish or circulate SET results in which individuals can be identified, if not by students, then by their professional peers. With these constraints in mind, it would seem far preferable for universities to use their carefully constructed lists of graduate attributes as a basis for assessing both student progress and teaching strategies. This would remove student opinion from the arena of customer satisfaction survey and could become a useful reference point for both students and staff. It would allow comparisons over time and across faculties so that some meaningful adjustment could be made to teaching strategies. No doubt universities would make some real gains from such monitoring, the broader public would be better informed about student use of public resources and the Commonwealth may at some stage be obliged to acknowledge that education cannot be compressed into the unsuitable mould provided by the service sector of the Australian economy. The authors thank those lecturers who made their student evaluation results available for this study.

Author Biography

Peter Slade is a Senior Lecturer in Economics and Finance at the University of the Sunshine Coast. His research interests have included the economics of crime, labour economics, industrial relations, house pricing dynamics and tourism. Prior to becoming an academic he spent many years in forestry in Australia and New Zealand.

Chris McConville has had wide experience at a number of Australian universities and in the older College of Advanced Education sector. He has taught in a range of humanities areas as well as in a number of specialised professional areas including planning and other aspects of urban environmental management. Chris has worked outside universities for local government and in media.

References

Abrami, P.C., d'Apollonia, S., Cohen, P.A. (1990). The validity of students' ratings of instruction: What we know and what we do not. *Journal of Educational Psychology 82*, 219-31.

Adams, J.V. (1983). Student evaluations: The ratings game. *Inquiry, 1*(2), 10-16.

Aleamoni, L.M. (1987). Typical Faculty Concerns about Student Evaluation of Teaching. In L.M Aleamoni (Ed.), *Techniques for Evaluating and Improving Instruction* (pp. 25-31), New Directions for Teaching and Learning, No. 31. San Francisco: Jossey-Bass.

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

Cashin, W. (1990). Students do rate academic fields differently. In M. Theall & J. Franklin (Eds.), *Student Ratings of Instruction: Issues for Improving Practice, New Directions for Teaching and Learning,* No. 43. San Francisco: Jossey-Bass.

Caskin, W.E. (1983). Concerns about using students' ratings in community colleges. In A. Smith (Ed.), *Evaluating Faculty and Staff: New Directions for Community Colleges.* San Francisco: Jossey-Bass.

Cohen, P.A. (1990.) Bringing research into practice. In M. Theall & J. Franklin (Eds.), *Student Ratings of Instruction: Issues for Improving Practice* (pp. 123-132), New Directions for Teaching and Learning, No. 43. San Francisco: Jossey-Bass.

Damron, J.C. (1995). The three faces of teaching evaluation. *American Educational Research Journal, 23*(1), 17-28.

Damron, J.C. (1996). *Instructor personality and the politics of the classroom*, from www.mankato.msus.edu/dept/psych/Damron-politics

Deming, W.E. (1986). *Out of the Crisis*. Cambridge, Massachusetts: Massachusetts Institute of Technology Center for Advanced Engineering Study.

Deming, W.E. (1993). *The New Economics for Industry, Government, Education.* Cambridge, Massachusetts: Massachusetts Institute of Technology Center for Advanced Engineering Study.

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

Emery, C.R., Kramer, T.R., & Tian, R.G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education, 11*(1)*,* 37-47.

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, 24*, 139-213.

Feldman, K.A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R.P. Perry & J.C. Smart (Eds.), *Effective Teaching in Higher Education: Research and Practice,* 368-95. New York: Agathon Press.

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

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, C.B. (1998). *Multivariate data analysis,* (5th Ed.). New Jersey: Prentice Hall International.

Hunt, L.M. (1997). *Another option to deliver better results*. In Proceedings of the 15th Annual Symposium of the Royal Aeronautical Society*,* Palmerston North, New Zealand.

Isley, P. & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *Journal of Economic Education, 36*(1), 29-43.

Martin, J.R. (1998). Evaluating faculty based on student opinions: Problems, implications and recommendations from Deming's theory of management perspective. *Issues in Accounting Education, 13*(4), 1079-1095.

Reckers, M.P. (1996). Know thy customer. *Journal of Accounting Information,* Summer, 179-185.

Russell, B. (1976). The study of mathematics. In B. Russell, *A Free Man's Worship and Other Essays.* London: Unwin Paperbacks.

Theall, M. & Franklin, J. (1990). Student ratings of instruction: Issues for improving practice, Introductory Section. In Theall, M. & Franklin, J. (Eds.), *New Directions for Teaching and Learning,* No. 43. San Francisco: Jossey-Bass.