Using student evaluations to improve individual and department teaching qualities

Mary R Hedges[1] and Don J Webber[2]

[1] *Centre for Longitudinal Research, University of Auckland, New Zealand*
[2] *Department of Accounting, Economics and Finance, University of the West of England, Bristol, UK*

Abstract

Student evaluations can be seen as an opportunity for students to vent their views on the quality of teaching that they receive, and sometimes instructors trivialise the importance of this information exchange opportunity. This paper takes student evaluations of teaching quality seriously and highlights that the information can be used more effectively. It illustrates how information from evaluations can be used to identify areas where the whole department has strengths and weaknesses and where individual instructors perform relative to their own department. This information can be used to identify individuals with specific areas of expertise and shape best practice within departments, across departments and/or across institutions. It can also be used to highlight individuals who may require further training and reveal areas of mediocrity that are at risk of intervention from a higher level. Finally, it suggests ways to implement shared best practice in order to improve department teaching quality assessment results and individual teaching performance.

Keywords: Student evaluations; Buddy schemes; Centres of teaching excellence

JEL codes: A220

Acknowledgements: For very useful comments, the authors would like to thank seminar participants at the DEE conference held at LSE in 2011 and Andrew Mearman, Paul Latreille and Ian Brailsford.

Corresponding author: Mary Hedges, Centre for Longitudinal Research, University of Auckland, Auckland, New Zealand. Email: m.hedges@auckland.ac.nz

1. Introduction and Background

For as long as there have been systematic student evaluations of instructor[1] performance there have been debates regarding their validity, usefulness, appropriateness and even more heated debates as to the use of the results as a management tool. From this literature a list of commonly held perceptions toward student evaluations of instructors have arisen. A comprehensive summary of this literature, going back to 1924, is provided by Aleamoni (1999). This summary highlights the extensive use of student evaluations and focuses on: the ability of students to accurately and/or appropriately make judgements about their instructor's competency; the validity of the surveys themselves and; the role of external factors such as class size, time of day, instructor qualifications, course level etc. Most of the arguments, or myths, about student evaluations that are de-bunked by Aleamoni (1999) are those used by academics themselves in order to minimise the weight put on any evaluations and/or negate their use and value. In fact, the author concludes that

> "The disadvantages of gathering student ratings primarily result from how they are misinterpreted and misused. The most common misuse is to report raw numerical results and written comments assuming that the user is qualified to interpret such results validly. Without normative or comparative information, a faculty member will be tempted to place inappropriate emphasis on selected student responses. If the results are published by the student government association, the biases of the editor(s) might misrepresent the meaning of the ratings to both students and faculty. If administrators use the ratings for punitive purposes only, the faculty will find ways to undermine their use and impugn their credibility." (p160)

This paper instead suggests ways in which managers can use the evaluations for constructive purposes and use them as an adjunct to building a culture of team co-operation. Section two considers some of the aspects of student evaluations, in terms of the design of them, consistency across time and that they are only one form of lecturer evaluation. This highlights that they are most useful when combined with other forms, both quantitative and qualitative, of lecturer review and evaluation. Section three briefly examines variation in policy versus implementation in evaluations at two sample institutions. These differences have implications in terms of the willingness of instructors to allow them to be used as a part of constructive management processes to improve individual and departmental teaching quality. Section four presents examples of how comparisons between peer and department averages can provide a tool to enhance peer support and professional development to both the students' and instructors' benefit. Section five deliberates about some of the broader issues in attempting to use evaluations in this way and other pre-conditions that may need to exist in order for it to be a viable strategy. This includes discussion on the need to integrate instructor evaluation processes with wider moderation and research collaboration processes. Section six then draws conclusions.

2. Instructor Evaluations

In reviewing the literature on instructor evaluations there is substantial diversity but they are unified through the common themes of the complexity of teaching (and the academic context)

[1] Within the context of this paper the terms instructor, lecturer and teacher will be used inter-changeably. There is no intended difference between methods or philosophy used between the three terms.

and their role in critical reflection by individual teachers. It is worth briefly considering these two aspects.

There is little doubt that higher education is becoming more complex and the pressures in academia are increasing[2]. Some of this comes from increased measurement, or attempted measurement, of what are considered to be key performance indicators in academia. Experience is relatively easily recorded and many countries now have some form of research measurement[3]. The focus has now shifted among policy makers and administrators to how best to measure teaching. While student evaluations are but one part of any such evaluation exercise it is often the primary tool used as it is relatively low cost when compared to many alternatives such as teaching portfolios (Langbein, 2008). What makes any measurement in this area more complex is that the relationship between evaluations and reflective teaching is not necessarily two-way. While reflective teachers usually consider student evaluations (among others) in their reflective practice there is no need for evaluated teachers to necessarily be reflective ones (Brookfield, 1995).

Knight (2002a) discussed three aspects to getting good evaluations – students, colleagues and oneself – omitting the link to theory offered by Brookfield (Brookfield, 1995). Later in the book Knight links very tightly to both theory and the idea of the observer's tacit theories that underpin what they are doing and/or reviewing in a much more critically reflective tone. It is worth briefly considering each of the three aspects identified by Knight as they provide a useful conceptual link between the literature and practice. It also aids in explaining how evaluations fit with models of the reflective teacher.

This paper is about the use of student evaluations therefore it is the student who has primacy in providing this feedback; however, what are they providing feedback on? All instructors will be aware that a student's interpretation of what is done in class may be quite different from what is intended (Trigwell, 2003). Their perceptions and the reporting of the match between their expectations and experience are important background to what they will say on evaluations. The responses on these evaluations therefore provide the instructor with useful information as to how well they explain themselves and their teaching methods to their students. In some literature on how to get good evaluations it is about modifying students' expectations rather than necessarily improving performance – explaining why things are done rather than just what is to be done (Edmundson, 1997).

There is also a substantial literature[4] on the link between expected and/or actual grade and student evaluations of instructors. Much of this work looks at controlling for student, instructor, course and department differences and the impact of evaluation processes on grade inflation. This literature also mentions various ways that the final numerical score of evaluations can be adjusted for factors such as course difficulty, class size, composition of the class in terms of majors/minors or elective takers and level (Ewing, 2012). Where this type of adjustment is being undertaken student evaluations are already being used as a key management tool where some form of comparability between instructors and/or departments is necessary. This type of adjustment is not undertaken at either of the universities discussed here. Instead, the approach taken here is focussed on using these evaluations to build

[2] See for example (Bond & Paterson, 2005; Forgasz & Leder, 2006; Glassick, Huber, & Maeroff, 1997; Harris, Thiele, & Currie, 1998; Jones, Galvin, & Woodhouse, 1999; Kearns & Gardiner, 2007; Kirp, 2003; Knight, 2002b; Macfarlane, 2005; Malcolm & Tarling, 2007; Sparkes, 2007; Thompson, Constantineau, & Fallis, 2010).

[3] Research Assessment Exercise (RAE) in the UK, Performance Based Research Funding (PBRF) in New Zealand and the Excellence in Research in Australia (ERA) in Australia.

[4] See for example (Hamermesh & Parker, 2005; Isely & Singh, 2005; Johnson, 2003; Langbein, 2008; McPherson, 2006; Nelson & Lynch, 1984; Rojstaczer & Healy, 2010; Sabot & Wakeman-Linn, 1991)

collegiality rather than as a basis for enrolment management and/or promotion or remuneration decisions.

There are a number of measurement and incentive issues with the use of student evaluations. The little meta-analysis undertaken on their consistency and validity suggests that well-developed instruments and procedures can provide high internal consistency reliabilities (ranging from 0.81 – 0.98 from the studies cited in Aleamoni, 1999). Unfortunately in less well-designed and administered evaluations, as is the case with many student- and faculty-generated forms, the reliabilities may be so low as to negate the evaluation process completely (Everly & Aleamoni, 1972). The focus of this paper is on the use of professionally developed evaluation systems that are university-wide and both the universities used as examples in this work have a very high consistency between the questions asked and scale used on their respective evaluation forms.

When it comes to comparing the validity of student evaluations with other indicators of instructor effectiveness there has been considerably less analysis undertaken. However, that limited body of evidence suggests a high degree of validity with other well designed evaluation systems, particularly those that also include qualitative components (Abrami, d'Apollonia, & Cohen, 1990; Costin, Greenough, & Menges, 1971; Koon & Murray, 1995; Marsh, 1984). Supporting this validity literature other studies have found high correlations between student evaluations over time (Hativa, 1996; Hogan, 1973; Marsh & Overall, 1979) and other indicators such as peer (colleague) ratings, self-ratings, expert judges' ratings (Aleamoni & Yimer, 1973), graduating seniors' and alumni ratings (Drucker & Remmers, 1951) and a more recent stream of literature that relates them to student learning (Davies & Guest, 2010; Doyle, 1975; Mirus, 1973).

Based on this range of literature the authors of this paper agree with the general validity of well-constructed student evaluations. Many studies already cited find these evaluations provide highly consistent results across time for a single instructor (if the evaluation questions used are not changed). Hogan (1973) goes so far in his analysis as to determine that approximately six sets of evaluations are needed in order to establish a reliability of approximately 0.90. These consistency findings are not a surprise when we compare these evaluations with consistency in student performance. Students usually get similar grades between assessment items within a course and grades across their courses. There may be outliers for any number of special circumstances (as may occur with lecturer evaluations) but a pattern of performance quickly becomes evident. For these reasons this paper will not go further into discussion on the validity, reliability and consistency of evaluations but accept that for well-designed evaluations, that are being completed anyway, there is additional value that can be gained from using them.

With regards to colleagues, Costin (1971) brings out a number of useful points such as honesty, responsiveness, relevance, respect, openness and empowerment but also explicitly links these to the idea that several observations are required before it is possible to claim the judgements as reliable (compare this with Hogan's findings). Aleamoni (1973) also discussed those aspects as being important for teaching though he did not make them equally important in the reflection and 'conversation with colleagues' stage. This is something that does not appear to be a core component of how these evaluations are run at many universities and this is discussed in more detail in the next section. An additional finding by Aleamoni (1973) is that instructor rank plays an important role in colleague evaluations and this is given more weighting than actual teaching performance. This finding explains why the authors found greater variation between colleague ratings of instructors than between student evaluations but they also raise a number of confounding variable possibilities.

The third critical aspect that Knight refers to is 'oneself'. This is essentially what a person does with their evaluations. This relates directly to the earlier quote from Aleamoni (1999) where he spoke of the 'inappropriate emphasis' faculty may place on evaluation results, both quantitative and qualitative aspects, if there is no normative component included. For example, it is very easy to scan through the results, decide they look 'normal' and file them; this is not the behaviour of a reflective teacher. It is also possible to read through them and place such inappropriate emphasis on all the negative/positive comments and then stress the next time you go to class that these students won't/will like you either/too; that is not the behaviour of a reflective teacher.

Reflective practice involves reading the feedback, identifying areas of performance where there may have been improvements or deterioration, relating these to the normative or comparative information, to any special circumstances of that particular class (and this may be external to the instructor such as location, time of day, etc. as analysed in more detail by (Shapiro, 1990), your personal philosophy and approach to that class and considering changes that you could make in future to alter these type of responses. For many reading this paper it would seem that the last of these options is the obvious approach and the question is probably being asked – isn't that what everyone does? The answer is – no. The 'why not?' is more interesting and will be considered in the following sections of this paper.

3. University Policies and Implementation

Discussion between the authors and viewing the student evaluation policies of the two universities we were employed by when this paper was initiated (and drawing on previous institutions as well) it was found there was substantial variation between what is reported and what is actually done at different institutions. When reading the rhetoric and/or policies at the two universities, they appear to be almost carbon copies. When these are then compared to the reality, usefulness and constructiveness of the two processes they are not in the same galaxy.

At one institution (that differentiates itself on its quality teaching) the management focus was on the pattern of responses over time and across courses. Furthermore, for promotion and professional development purposes it was how the teacher had critically evaluated and/or responded to these evaluations that was considered rather than the results themselves. The actual processes involved included that all student evaluations must be conducted by an academic colleague and when the processed results arrive the two parties must meet to discuss them and consider any special circumstances, issues etc. or possible ways forward. The summary of this discussion is then documented on the report and must be signed by both parties. Over time managers would expect to see some efforts or signs of progress in the areas of identified weaknesses. This could be as simple as a reduction in the number of comments that refer to something. For example, if early evaluations included a number of comments about talking while students were expected to be taking something down and this disappeared in later evaluations, then that would be considered progress.

The second aspect of this process was that it developed a level of trust and respect between peers as they had to undertake each other's evaluations and discuss them. This part of the process meant that they were not only student evaluations but also incorporated peer evaluation components and facilitated the discussion of issues, weaknesses and successes with one's peers. This level of trust will become important in the next section of this paper.

This implementation process is in stark contrast to the other institution (that prides itself on its international research ranking) where the evaluations are often undertaken by administrative staff or tutors who simply process them and play no further part in the process.

There is no requirement to discuss them with a peer and only minimal requirements to even complete them on a regular cycle. Even if completed there is no requirement to include all of them in any promotion application. While copies are sent to Heads of Department they seldom become the focus of any professional development of discussion unless there is a clear, and usually significant problem.

Given this variation in the implementation of what appear to be very similar policies there will clearly be limits in the applicability of the rest of this paper. However, perhaps we can provide aspirational insights into the potential uses of evaluations to improve teaching quality.

4. Using Student Evaluations to Improve Teaching Quality

Questions used on well-designed student evaluations are usually consistent within institution, faculty or department. They provide a rich data source that could be used to improve teaching quality; however, the data are frequently largely ignored in this context, often due to fears about privacy and/or competition. We suggest that these evaluations can be used constructively in order to build department reputations and explicitly to improve overall department teaching quality. This section will outline this process.

Feedback scores

Student evaluations of instructors can be used to identify what students think instructors and departments are relatively good at, and relatively poor at. This can be calculated by collating the responses from all evaluations on all modules taken by individual instructors and from an entire department. Care does need to be taken in doing this so that it is collated at the end of an evaluation period when all responses have been collected. If collated cumulatively through the collection period it can penalise the relative position of courses depending on the order in which the results are added to the sample[5]. Care also needs to be taken when a module does not run, or when instructors are new to a course which can lead to substantial variance depending on the modules being offered and other external factors over which instructors have no control. For example, if all theory courses are offered in one semester and survey/elective courses in another semester then this could lead to significantly different results for the two semesters so comparing like-with-like would become an important consideration.

By averaging the evaluation results across all modules[6] a ‘peer average’ can be constructed; potential issues related to the calculation of a peer average will be discussed in the final section and relates to the literature on the numerical adjustment of evaluations (Ewing, 2012). If the peer average is taken at faculty level then it is possible to identify particular departments’ relative strengths and/or weaknesses. If the peer average is taken at department level it is then possible to identify particular individuals’ relative strengths and/or weaknesses. For the rest of this paper the discussion will proceed as if it is a department peer

[5] Some universities even publish the ‘Department average’ or ‘Faculty average’ though this is often subject to timing issues of when particular courses are evaluated. One of the universities here no longer publishes this information for that reason.

[6] The authors would recommend that a good rule of thumb for how many to include is to group them in such a way that a minimum of six evaluations (Hogan, 1973) of each lecturer and course be included in any aggregate measure. Large departments could then split according to the level of the course while smaller department would tend to keep them together.

average and an individual instructor but it is easy to see how the process could be scaled up to Faculty/department level.

It is then possible to compare these peer average values against individual instructor's values. This will illustrate where instructors are good at something, and a Head of Department can use this information to encourage them to guide other members of the department and build teaching teams based on skill mixes when teaching teams are required. Most of us would be familiar with the underlying rational here: as instructors we often put students into groups for group work based on balancing grade, gender, major or other characteristics. Why not build teaching teams the same way? Teaching teams do deliver positive results in terms of course continuity and managing operational constraints but they can also be problematic if not created and managed effectively. This in turn relates to the degree of collegiality and collaboration in a department that will be discussed in section 5.

What is important is that all of these values/measures need to be *relative* to the peer average at either department or faculty level. To permit such an analysis there needs to be some harmonisation or standardisation of questions on the evaluation sheet and this again requires that they be well-designed evaluations designed by experts. Some standard evaluation questions are listed below. These are examples of questions that relate to the numbers on the x-axis in the following diagrams:

1. The lecturer is approachable
2. The lecturer is organised and well prepared
3. The lecturer's enthusiasm helps me to learn
4. The lecturer helps me learn by using useful explanations and practical examples
5. The lecturer seeks and responds to feedback from students
6. Overall the lecturer is a highly effective teacher

Examples of other questions commonly used include:

- The lecturer effectively uses his/her subject knowledge to guide my learning
- The lecturer assesses my understanding when teaching and gives me constructive feedback about my progress
- The lecturer clearly communicates assessment requirements
- The lecturer treats me with respect
- The lecturer communicates effectively
- The lecturer creates a positive learning environment for me
- The lecturer helps me to take responsibility for my own learning.

Figure 1 illustrates a hypothetical example to illustrate the ability to use this information as an important source for constructive continual professional development. The x-axis lists the questions sourced from the student evaluation questionnaire while the y-axis presents the score for that question. The department is represented by the circle and the individual instructor by the diamond. The line shows the degree of variance of the department average and the line extends one standard deviation in each direction from the department average. While technically it would be possible to calculate the department average excluding each staff member in turn we see little benefit in this based on the previous discussion. Instead we recommend a standard department average, in which each individual will have some/all of their personal evaluations included. This means that the magnitude of the difference for any individual from the department average are biased in such a way that the individual to department difference computed (distance between the diamond to dot respectively) will be less than actually exists.

{Insert Figure 1 about here}

These hypothtical results have been arranged in decending order such that the best performing trait of the department comes first. They have also been ordered relative to an average across all questions in order to identify relative strengths and relative weaknesses. The variance line illustrates the degree of variability in the responses and show plus or minus one standard deviation from the department aveerage. From this we can see that the responses to questions 1, 4 and 5 for this instructor were not different from the department average. We can also see that the department responses to question 4 were particularly varied (longer variance line). The aims of a department may be two-fold: to increase the average quality of teaching and; to reduce the variation in teaching quality. Based on this example this suggests different areas to focus on in response to each of these objectives.

1) The department should focus on reducing the range of teaching quality for area 4, the use of explanations and practical examples. For instance, the department could adopt a policy of necessarily integrating at least five high-quality examples per teaching hour. A regular opportunity for staff to share the examples they use could be introduced where there is already a degree of collegiality. In environments where this type of interaction is uncommon the use of an external moderator in the first instance who can provide a baseline and stimulate discusison may be one approach to introducing a new ‘teaching’ element into department meetings.

   It also looks as if area 6 has a high level of variance but as this score is an ‘Overall’ measure it is possible that by reducing the variance in area 4 this will have a flow-on effect to area 6. The links between components such as this is considered by Balemi (2012) where he takes an indivdiual approach to how the component questions feed into the overall score.
2) Lecturer X is within one standard deviation of the average for areas 1, 4 and 5. Although lecturer X should be congratulated, with the rest on the department, in achieving a good score in area 1 (being approachable) the whole department, including Lecturer X, should also be encouraged to do better in area 5 (seeking and responding to feedback). A policy response to this may be that a feedback sheet should be revised to enable the marker to provide more detailed feedback on specific areas. Alternatively it could be that substantial feedback is already provided but the students don’t recognise it as feedback. In this case better explanations could be provided that make explicit what feedback is.
3) Lecturer X requires further training in areas 2 and 3 (organisation and enthusiasm respectively). Two policy responses could be the result. First, for area 2 (organisation) this instructor is close to the department average. While positive this whole department could improve in this area. Perhaps having a department in service workshop facilitated by an external (to the department) expert in this area could assist everyone. Second, the relatively low level of enthusiasm for the topic could be related to this course being tangental to the instructors areas of expertise. If this is the case it could be addressed by guiding the individual to teach a different module that is closer to their personal interests or research area or agreeing to some trade-off in order to meet operational constraints.

It is no acident that these two areas are together. Research by Costin (1971) and Crawford (1968) found that the four most mentioned and important characteristics of highly rated instructors were: *(a)* thorough knowledge of subject matter, *(b)* well planned and organised lectures, *(c)* enthusiastic, energetic and lively teaching, and *(d)* student oriented. The order of these is interesting and potentially also suggests causality – when you know the subject matter well you can be better prepared and it is easier to be more enthusiastic.

4) In area 6 (overall effectiveness) the department is doing relatively poorly, although there is a large range of performances within the department. In this example Lecturer X does particularly well in comparison to the department. This suggests that this lecturer could lead the way by sharing best practice with his/her colleagues in order to improve their performances and reduce variation. This could be done in two ways: internal best practice sharing at department level or a more informal 'buddy' system. We use the term 'buddy' here to refer to an informal collaboration between peers where they can share and support one another with a focus on specific areas. For example a young, dynamic, innovative teacher could be paired with a cynical, older research specialist. They both bring specific skills and enthusiasm to the partnership and through collabortation on both fronts they can both improve their academic credentials. It is not intended that such a system be a part of management systems although it may be the manager that suggests the staff work together. Often it is this score that most emphasis is placed on by both indivduals and managers so undertanding the link between component questions and this question is important (Balemi, 2012).

In this way it is possible to use student evaluations to identify individual strengths where leadership roles within the department/faculty could be developed. It also identifies areas where efforts within the department could be focussed in order to improve teaching performance, either for the individual as in areas 2 and 3 or for the whole department as in areas 4-6. This teaching support could either be internal to the department, as in area 6, or external to that department, as in area 5. Area 4 would be an interesting one where gains could be achieved both through internal buddying and also by calling on external professional development opportunities.

Quadrant analysis

Another way of looking at student evaluation feedback is through quadrant analysis where the individual is mapped on the y-axis and the department (or peer average) is mapped on the x-axis. Each point therefore represents the individual in relation to their department but again, because the individual's evaluations also contribute to the department averages, where there is a difference these will be biased in such a way that the plotted difference will be less than the real difference. In this example each point has been numbered to match the questions used in the hypothetical example above for ease of discussion. If you take question 1, the department average from the figure 1 was 0.2 and the individual was 0.1. These are now plotted as a single point located (y=0.2, x=0.1*)*. Variance is now omitted but instead each quadrant has a direct interpretation to guide supportive continual professional development initiatives.

{Insert Figure 2 about here}

Quadrant A: The top-right quadrant is where the department is doing well at something as is the individual staff member. This is the "keep up the good work" quadrant. It represents the area(s) in which the department is doing relatively well on its own student evaluations and could be an area which the department could provide support to other departments within and outside of the university. Lecturer X is receiving good feedback on question 1 but when this information is used in conjunction with Figure 1 it can be identified that most other department members could also be at least as good as Lecturer X in this area. Lecturer X's average score across all questions is equal to his score for question 2; this is an area which the department is also performing relatively well. If when comparing department to faculty performance this department still scores highly in this area a case could be made that this department be a centre of excellence in teaching in this particular area. In addition the department could stake a public claim that they are performing well in this area. Similarly, if students want to receive a teaching experience which is characterised by a particular teaching quality then this information could also improve the student-department match during the recruitment process.

Quadrant B: Bottom-right quadrant is where the department is doing well at something but Lecturer X should be guided to concentrate efforts on improving their own performance. The department head could encourage this individual to buddy with a departmental colleague who is good in that specific area. The involvement of the department head is due to their access to this information on each staff member in the department and not for any punitive management reason.

Quadrant C: The top-left quadrant is where the department is doing relative poorly, but the individual staff member is doing well. Lecturer X should be encouraged to share best practice and is an obvious choice to act as buddy to their colleagues in this area. Sharing knowledge and experiences in order to improve student questionnaire scores will improve the impression of the department's teaching, and this will be illustrated in national ranking as an improvement in student evaluation performance.

Quadrant D: The bottom-left quadrant is where the department average and Lecturer X are both doing relatively poorly. The whole department should focus their efforts here, perhaps by obtaining training from other departments within the institution or from outside via a recognised centre of excellence. This trait would be an ideal candidate for a department professional development day.

5. Discussion

These approaches to utilising the rich data already provided by student evaluations suggest an obvious management approach, however, in addition to having strengths and weaknesses, it is one that would likely meet substantial resistance in some environments. Part of this comes back to the two different approaches taken in the implementation of an evaluation policy already mentioned and the 'chicken or egg?' relationship between collegiality and trust among staff within a department or a faculty.

In section three it was emphasised that the actual implementation of very similar policies was quite different at the institutions that the authors of this paper worked at. However, how widespread and representative of institutional practice these personal experiences are is not clear and is clearly an avenue for future research. In part this reflects the high level policy versus the management interpretation of that policy. However, another component is normative in nature and stems from how student evaluations are valued and treated within faculties and departments. This often comes down to two key pre-cursors for the preferred 'peer review' model: the emphasis put on quality teaching by management on a daily basis and; the level of peer support, team teaching and collegiality already present in any department.

The teaching focussed institution referred to here has a formal structure of formal mentors for new staff where new staff have a single, more senior and experienced colleague to whom they can take questions and concerns to. In addition there is a strong moderation policy and process that requires comprehensive involvement throughout the course as every assessment item undergoes detailed moderation by the same moderator. This is to ensure that overall assessment weighting and focus is aligned with the learning objectives and teaching time allocated to topics. Through these two processes all staff are familiar with courses taught by their colleagues. This creates a culture where there are incentives for staff to work collegially, to support one another and to share problems and successes. Within this environment, having colleagues run the evaluations and having to discuss them was a natural and constructive process that did enable integration of aspects of both student and peer evaluation.

In the second institution academic freedom is the dominant underlying philosophy and this has led to a culture where peer review is reserved for research rather than teaching. In some instances where peer review on teaching has been undertaken, as a separate process from student evaluations, it has been found to improve relationships and understanding. However, this was elective peer review between people with similar interests and it is questionable whether the same positive results would be gained if undertaken as part of a change in policy and a top down directive. Secondly, peers chosen tended to be those with a specific interest in critically reflective teaching quality so there was a starting presumption that they would be good teachers explicitly seeking constructive feedback. Would this potential value be recognised by those who are not such reflective teachers?

The key component to this proposed strategy is that of benchmarking teaching performance against an appropriate measure. In this context what could be more appropriate than a measure that the individual is a part of such as a department score. How to construct such a peer average has been controversial historically in at least one of the institutions discussed here. In most cases a stable average is gained if the last three years of evaluations were used in a rolling average manner. If instructors teach a course once per semester this is equivalent to six evaluations. This overcomes some of the potential issues related to staff and course changes[7]. Depending on the size of the department it could even be possible to have benchmarks relevant to different levels such as a first year benchmark and an upper-undergraduate benchmark. This could, for large departments, overcome some of the issues that it is easier to get good feedback from smaller, higher level papers than from large impersonal first year courses.

A major issue when using this type of analysis to act as a foundation for continuing professional development policies[8] is that the comparator modules are relevant. It is likely that claims could be made that particular modules could receive evaluations that are less likely to receive high grades; examples may include mathematical modules such as econometrics. It may also be a possibility that servicing modules (such as an economics module being taught to modern language students) may receive idiosyncratic feedback: they

[7] You need any peer average to represent your current staff so if there is high staff turnover you may need to run a shorter time period. Similarly is there is significant variation in the types of courses offered across semesters and/or years a shorter peer average period may be appropriate.

[8] There could equally be a case made for individuals who are excelling at a particular trait to also engage in continuing professional development in that area in order for them to provide better quality feedback to a department. For instance, Lecturer X may be performing relatively well at area 6 but they may not know why this is the case.

may be compared to the language students' other language modules rather than to other economics modules. It is also possible that there may be inherent sample selection issues whereby more able students naturally cluster into particular modules perhaps more technical in nature (such as econometrics) and less able students naturally cluster into other particular modules (the economics of culture and sport?). A further set of issues may occur when students perceive the jump between years to be too great or not great enough; for instance, students may find their first year at university to be too hard or too easy depending on their education background. Associated with this is whether student evaluation university league tables are used to such an extent that making the students happy rather than more educated becomes the prime, though informal, objective.

Our recommendation is that departments should consider the use of this type of analysis to act as a framework on which to build improvements in teaching quality. We suggest that it is used as a formative guide in order to improve individuals and departments' performances. We do not recommend the use of this type of analysis as a summative process due to the caveats raised in the above paragraphs nor should it be included directly in any promotion/remuneration process though the indirect benefits of being a mentor or buddy or improved evaluation results obviously should be.

Lastly these types of scores can also lead to perverse incentives. For example, just as student evaluations can be improved by modifying their expectations, staff performance could potentially appear to be improved by lowering the quality of teaching in the department or faculty. That you are reading this paper we will assume that the goal is improving teaching quality – not just evaluation results.

6. Conclusion

There are a number of issues at every level of the proposed use of student evaluations to improve teaching quality and as a tool in professional development. Most of them do have solutions when well thought out and planned for. This is not very different from the improvements made in the student evaluations themselves over time as the quality, validity and consistency of these has improved. This paper has proposed a method of using the information obtained through student evaluations in a more effective way. The authors accept that the departmental and faculty contexts are a significant tool or barrier to the use of these evaluations in a structured and systematic way to improve teaching quality. However, we also believe that there is a 'chicken or egg' aspect toward building peer trust and support mechanisms. Explicitly using this information to build teaching teams could be one way to positively construe the introduction of such a scheme because of its direct link to classroom practice. The authors believe that the use of mentors, buddies and peer evaluations are also important tools in the critically reflective teacher's toolbox and if these can be effectively combined with the student perspective, gained through the evaluation responses, it can only lead to teaching quality improvement to the advantage of everyone.

Bibliography

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of Student Rating of Instruction: What We Know and What We Do Not. *Journal of Educational Psychology, 82*(2), 219-231.

Aleamoni, L. M. (1999). Student Rating Myths Versus Research Facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153-166.

Aleamoni, L. M., & Yimer, M. (1973). An Investigation of the Relationship Between Colleague Rating, Student Rating, Research Productivity and Academic Rank in Rating Instructional Effectiveness. *Journal of Educational Psychology, 64*(3), 274-277.

Balemi, A. (2012). *Better use of teaching/course evaluations.* Certificate of Academic Practice Research Projects. Reserach Project. Centre for Academic Development. University of Auckland. Auckland.

Bond, R., & Paterson, L. (2005). Coming down from the ivory tower? Academics' civic and economic engagement with the community. *Oxford Review of Education, 31*(3), 331-351.

Brookfield, S. D. (1995). *Becoming a critically reflective teacher*. San Francisco: Jossey-Bass.

Costin, F., Greenough, W. T., & Menges, R. (1971). Student Ratings of College Teaching: Reliability, Validity, and Usefulness. *Review of Educational Research, 41*(5), 511-535.

Crawford, P. L., & Bradshaw, H. L. (1968). Peception of characteristics of effective university teachers: A scaling analysis. *Educational and Psychological Measurement, 28*(4), 1079-2085.

Davies, P., & Guest, R. (2010). What effect do we really have on students' understanding and attitudes? How do we know? [Editorial]. *International Review of Economics Education, 9*(1).

Doyle, K. O. (1975). *Student Evaluation of Instruction*. Lexington, MA: Lexington Books.

Drucker, A. J., & Remmers, H. H. (1951). Do Alumni and Students Differ in Their Attitudes Toward Instructors? *Journal of Educational Psychology, 42*(3), 129-143.

Edmundson, M. (1997). On the Uses of a Liberal Education. *Harper's Magazine, September*(295)*,* 11.

Everly, J. C., & Aleamoni, L. M. (1972). The rise and fall of the advisor: Students attempt to evaluate their instructors. *Journal of the National Association of Colleges and teachers of Agriculture, 16*(2), 43-45.

Ewing, A. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review, 31*(1), 141-154.

Forgasz, H., & Leder, G. (2006). Academic Life: Monitoring Work Patterns and Daily Activities. *Austraian Educational Researcher, 33*(1), 1-22.

Glassick, C. E., Huber, M. T., & Maeroff, G. I. (1997). Scholarship Assessed: Evaluation and the Professoriate (pp. 130). San Francisco: Carnegie Foundation for the Advancement of Teaching.

Hamermesh, D. S., & Parker, A. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*, 369–376.

Harris, P., Thiele, B., & Currie, J. (1998). Success, Gender and Academic Voices. Consuming Passion or Selling the Soul? *Gender and Educaiton, 10*(2), 133-148.

Hativa, N. (1996). University Instructors' Ratings Profiles: Stability Over Time, and Disciplinary Differences. *Research in Higher Education, 37*(3), 341-365.

Hogan, T. P. (1973). Similarity of Student Ratings Across Instructors, Courses and Time. *Research in Higher Education, 1*(2), 149-154.

Isely, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *Journal of Economic Education, 36*(1), 29–42.

Johnson, V. E. (2003). *Grade inflation: A way out*: Springer.

Jones, G., Galvin, K., & Woodhouse, D. (1999). Universities as Critic and Conscience of Society: Th Role of Academic Freedom *AAU Series on Quality* (pp. 27). Wellington: New Zealand Universities cademic Audit Unit.

Kearns, H., & Gardiner, M. (2007). Is it time well spent? The relationship between time management behaviours, perceived effectiveness and work-related morale and distress in a university context. *Higher Education Research and Development, 26*(2), 235-247.

Kirp, D. (Ed.). (2003). *Shakespeare, Einstein, and the Bottom Line*. Cambridge, MA: Harvard University Press.

Knight, P. T. (2002a). Getting good evaluations. In H. Eggins (Ed.), *Being a teacher in higher education* (pp. 178-186). Buckingham: Oxford University Press.

Knight, P. T. (Ed.). (2002b). *Being a Teacher in Higher Education*. Buckingham, UK: SRHE & Open University press.

Koon, J., & Murray, H. G. (1995). Using Multiple Outocmes to Validate Student Ratings of Overall Teacher Effectiveness. *Journal of Higher Education, 66*(1), 61-81.

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review, 27*(4), 417-428.

Macfarlane, B. (2005). The Disengaged Academic: the Retreat from Citizenship. *HIgher Educaiton Quarterly, 59*(4), 296-312.

Malcolm, W., & Tarling, N. (2007). *Crisis of identity? : the mission and management of universities in New Zealand* Wellington, NZ: Dunmore Press.

Marsh, H. W. (1984). Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potentail Biases and Utility. *Journal of Educational Psychology, 76*(5), 707-754.

Marsh, H. W., & Overall, J. U. (1979). Long-Term Stability of Students' Evaluations: A note on Feldman's "Consistency and Variability Among College Students in Rating Their Teachers and Courses. *Research in Higher Education, 10*(2), 139-147.

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *Journal of Economic Education, 37*, 3–20.

Mirus, R. (1973). Some Implications of Student Evaluation of Teachers. *Journal of Economics Education, 5*(1), 3.

Nelson, J. P., & Lynch, K. A. (1984). Grade inflation, real income, simultaneity, and teaching evaluations. *Journal of Economic Education, 15*, 21–37.

Rojstaczer, S., & Healy, C. (2010). Grading in American colleges and universities. *Teachers College Record*.

Sabot, R., & Wakeman-Linn, J. (1991). Grade inflation and course choice. *Journal of Economic Perspectives, 5*(159-170).

Shapiro, E. G. (1990). Effect of instructor and class characteristics on students class evaluations. *Research in Higher Education, 31*(1), 135–148.

Sparkes, A. (2007). Embodiment, academics, and the audit culture: a story seeking consideration. *Qualitative Research, 7*(4), 521-550. doi: 10.1177/1468794107082306

Thompson, P., Constantineau, P., & Fallis, G. (2010). Academic Citizenship: An Academic Colleagues' Working Paper Retrieved 26 October, 2010, from http://www.queensu.ca/secretariat/senate/COU/AcadCitizen.pdf

Trigwell, K. (2003). A Relational Approach Model for Academic Development. In H. Eggins & R. MacDonald (Eds.), *The Scholarship of Academic Development* (pp. 23-33). Oxford: Oxford University Press and SRHE.

Figure 1: Average feedback scores

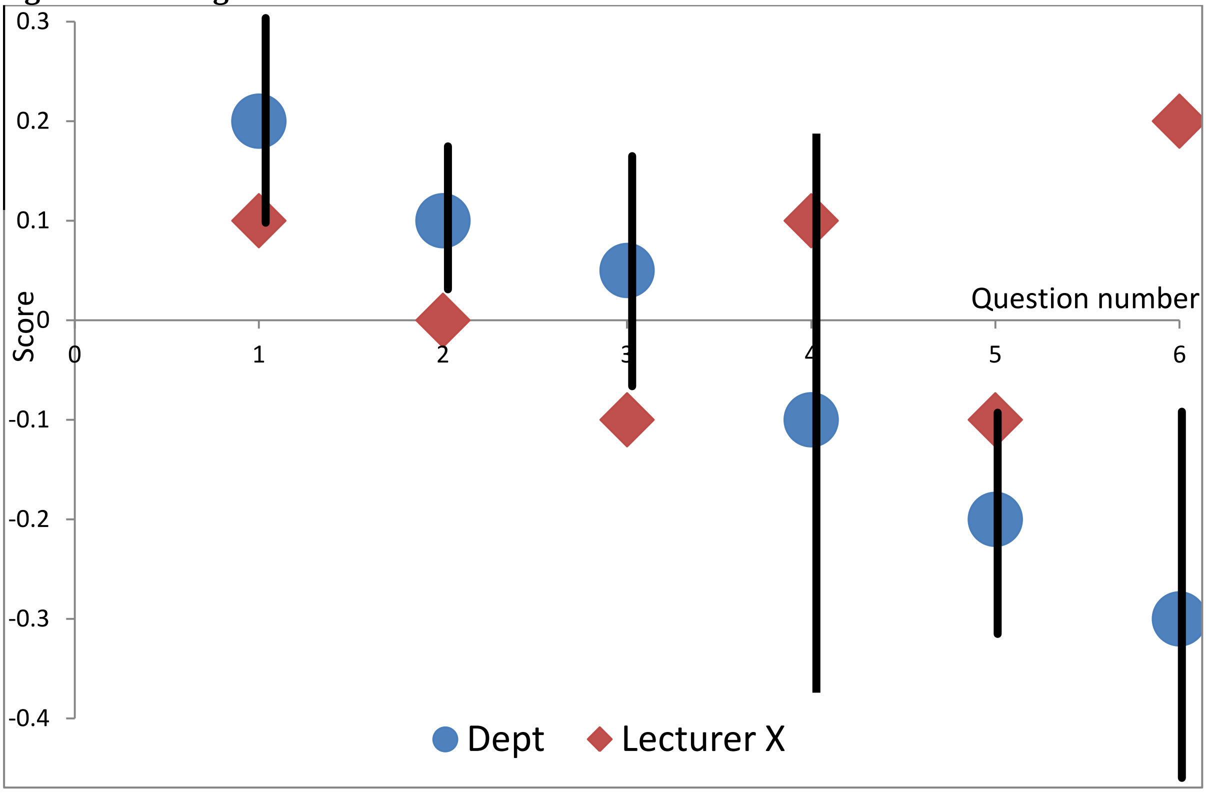
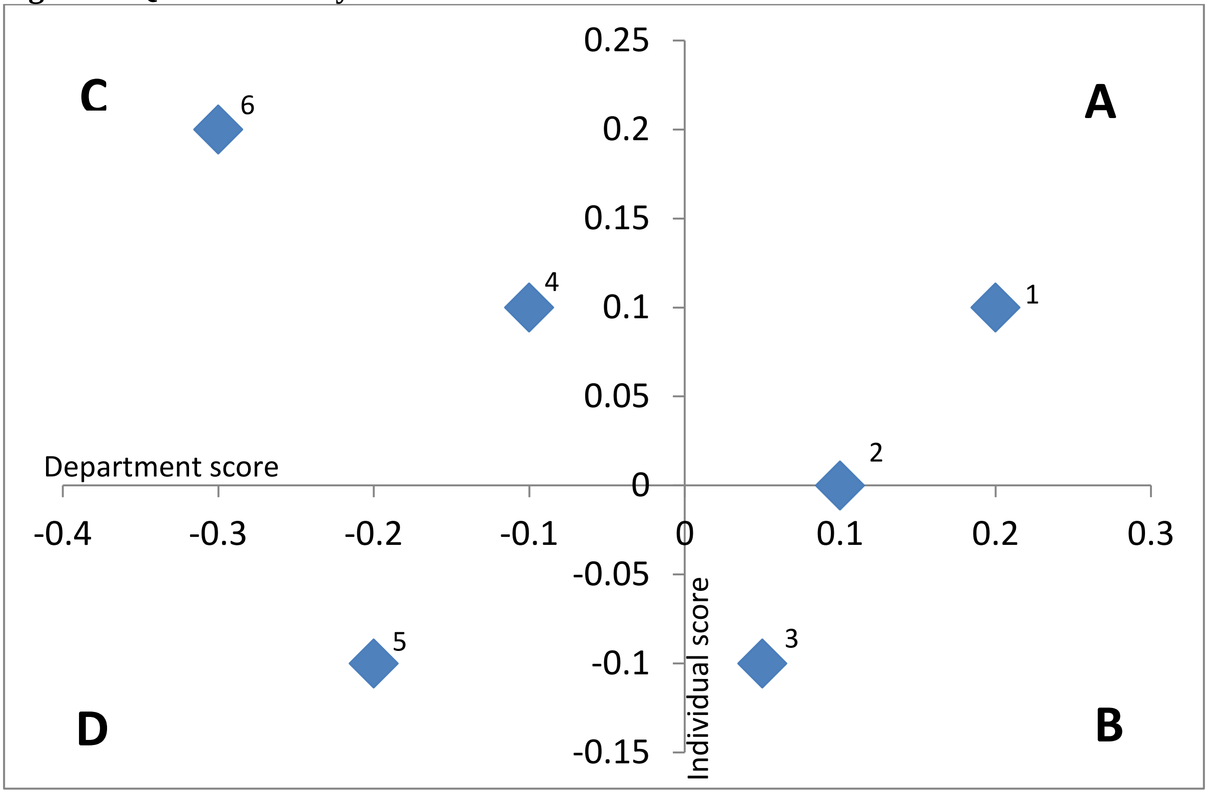


Figure 2: Quadrant analysis


C
A
D
B
Department score
Individual score