Is it your fault? Influences on student evaluations of teaching in tertiary institutions.

Martin Davies, Joe Hirschberg, Jenny Lye, Carol Johnston and Ian McDonald

Introduction

In Australia, as in Britain, there is an impetus to quantity the quality of teaching in order to be able to compare institutions and departments and in order to be able to compare individuals at the personal level. The use of surveys that collect information on student perceptions of teaching is now widespread. The Australian Commonwealth Department of Education Science and Training appears poised to use such survey data drawn from the Course Experience Questionnaire to rate institutions in Australia in relation to their teaching (DEST 2004). If the experience of Britain is an indication it is likely that Australian tertiary institutions will be ranked and funding will accrue to those institutions that rank highly. While it is not appropriate to debate here whether this is the best way to proceed in terms of improving tertiary education, it is appropriate to explore whether student perception survey data is robust enough to use for these purposes.

The widespread use of surveys of student perceptions of their teaching and learning experience has prompted an almost equally widespread investigation of what these surveys are actually measuring. There are literally thousands of references to research on student ratings of teaching, most in the last two decades. Most research has been conducted in the United States (see for review Cashin, 1995, and Marsh 1987. There are a handful of studies emanating from Australia (Bedggood & Pollard, 1999; Haynes, 2002; Marsh, 1987; Marsh & Bailey, 1993; Neumann, 2000; Wagner, 1999; Worthington, 2002) and Europe (Husbands, 1996, 1997; Husbands & Fosh, 1993; Shevlin, Banyard, Davies, & Griffiths, 2000). Few studies have been conducted over an extended period of time (Haynes, 2002; Marsh & Bailey, 1993; Marsh & Hocevar, 1991; Ting, 2000). The issue of the validity of student evaluations of teaching is a matter of some controversy with evidence both supporting their continued use and evidence recommending their discontinuation. There is no sign of the controversy abating (Abrami, d'Apollonia, & Cohen, 1990; Boice, n.d.; Cashin & Downey, 1992; Dwinell & Higbee, 1993; Greenwald, 1997; Hepworth & Oviatt, 1985; McKeachie, 1997; McKeachie & Lin, 1979; Smith, 2004; Solas, 1990).

The first part of this paper provides a broad overview of the existing literature in relation to student surveys and identifies potential biases and system influences. The second part of the paper reports an analysis of Australian student ratings in a department of economics collected over a ten-year period. Outcomes of this analysis and the implications for teaching practice are then discussed.

Purpose of collecting students rating data

Student rating surveys in tertiary institutions have been used to gain diagnostic feedback on teaching effectiveness (Marsh & Bailey, 1993; Marsh & Roche, 1993) in assessing the particular instructor characteristics that might assist in improving individual teaching performance. They have been used for administrative decision-making (McKeachie, 1997; Simpson & Siguaw, 2000; Wolfer & Johnson, 2003) for summative and formative purposes, as a basis for hiring, confirmation and promotion or in matching appropriate teachers to appropriate courses (Williams & Ceci, 1997; Wilson, 1998). They have also been administered to fulfil the demands for the accountability of the institution (Williams & Ceci, 1997) and so that cross institutional comparisons may be made

(Access Economics 2005). They have also been available in some institutions for students to use in informing their subject selection decisions (Marsh 1987). The appropriateness of using this rating data in these ways is a matter of deep controversy.

Sources of contention in relation to the use of ratings data

Several studies have confirmed the validity and reliability of student ratings in the USA (Bosshardt & Watts, 2001; Wachtel, 1998), in Australia (Marsh, 1987), in Hong Kong (Ting, 2000) and in Europe (Byrne & Flood, 2003; Husbands, 1996, 1997; Husbands & Fosh, 1993). Despite this there is a great deal of debate about the usefulness of these ratings. Some argue that the precise numerical scores generated in student evaluations 'imply a level of measurement that simply does not exist' (Wolfer & Johnson, 2003), p.117) and that it would be better if numerical scores were replaced with categories such as "exceptional", "adequate" and "unacceptable" (d'Apollonia & Abrami, 1997), more global measures of assessment (Cashin & Downey, 1992) or a range of scores (Neumann, 2000). Others suggest that the fine measurements generated do not allow discrimination between good and bad teaching and that "cutoff" points would be a more sensible approach (McKeachie, 1997) Still others claim that the use of student evaluations for administrative purposes is a misappropriation of the data (Bedggood & Pollard, 1999; Sheehan, 1975) in part as students themselves may have little idea that such use is made of their evaluations (Dwinell & Higbee, 1993).

A common form of instrument has been the student questionnaire that uses Likert-type scales as a measure of student opinion. Often, the ratings from such instruments are accepted without question, even though it is not known whether all students interpreted the items on such questionnaires in a similar way, or whether students were consistent in their interpretation of the rating scale. Students may reinterpret the meaning of the items in ways which may be quite different from the intentions of the questionnaire designer or researcher (Low, 1999). Block (1998), complemented the administration of a student questionnaire with a small number of student interviews and found a high degree of variance in the responses of students to questionnaire items and to the ratings they had given to those items.

The method of administration is also an area of debate and one that can affect the ratings given by students. Dwinell and Higbee (1993) found that a majority of students believed their anonymity was assured when they provided ratings of their instructors, while a Canadian study (Fries & McNinch, 2003) revealed that the anonymity had some effect on students' ratings of teachers. Students who had signed their ratings forms gave higher ratings to their teachers.

Kolitch and Dean (1999) point out that assumptions about curriculum, instruction, evaluation and student-teacher relationships underlie the items of a typical rating instrument. Even though a rating instrument might claim to be representative of all conceptions of teaching, such an instrument is more consistent with a transmission paradigm of teaching. Solas (1990) recognised that data obtained through surveys are confined to a small number of dimensions, often "more meaningful to the surveyor than the individual being surveyed" (p.152). Instead Solas used a repertory grid which enabled students to express their own opinions, in their own terms, thereby providing more detailed information than possible through a questionnaire (Solas, 1990). Other sources of data, such as student journals (Wagner, 1999) and teacher interviews (Ballantyne, Bain, & Packer, 1999; Hativa & Raviv, 1993; McCormick, 1996) have also been employed to gather student perceptions of their teachers.

The usefulness of student ratings in improving teaching is questioned. Students do not have the knowledge necessary for appropriate rating of teaching (Simpson & Siguaw, 2000). They are not necessarily best judges of their instructors' performance (Casey, Gentile, & Bigger, 1997), viewing it 'from very limited or even tainted perspectives' (p. 472). They consider the provision of feedback

as a chore (Simpson & Siguaw, 2000). For their part, though, students perceive that their evaluations of instructors are heeded and that instructors do change their behaviour in consequence (Dwinell & Higbee, 1993), even though later cohorts of students have no means of comparison, and so are unlikely to be aware of such changes (Haynes, 2002). Indeed, information gathered from such evaluations is of little use to rectify problems for current classes, and may be inappropriate for subsequent classes if the evaluations are conducted at the end of a term or semester (McKeachie, 1997). The usefulness of student evaluations can be assured only if data gathering, reporting and interpretation are carried out in a careful, considered manner (Casey et al., 1997).

The influence of different teaching contexts should be considered (Neumann, 2000) and scores generated should be weighted according to discipline area, class size and other factors (Wolfer & Johnson, 2003).

Influences on student rating data

The majority of studies have gathered quantitative data from rating surveys that use Likert-type rating scales. There are a number of factors that are explored in terms of their potential affect on the responses to these rating surveys. These can be broadly groups in terms of the influence teacher or instructor-level determinants, student-level determinants and subject-level determinants. Variables in the teacher-level category include the instructors' use of class time, their availability outside class time, how well they assess student learning or understanding, their concern for students' welfare and performance, the extent to which they emphasise analytical or critical skills, their preparedness, their tolerance of alternative viewpoints in class. Student-level determinants include the reasons for taking the course, the class-level of the respondent, the effort students expend in the subject, age of the student, ethnicity and student gender. Subject-level determinants include when the subject is offered, whether it is required for the degree or a prerequisite subject, the level, the perceived difficulty, grade inflation and the size of the class.

How do these factors influence student ratings of their teachers, and which factors are the most crucial? Are the measures biased or do they represent an accurate indication of instructor teaching effectiveness? Finally, what is the relationship between the student measures of teachers' performance and the teachers' measures of student performance? Can good instructor ratings, for example, be "bought" by teachers giving students good grades? Studies are reviewed in the following section in terms of instructor-level determinants, student-level determinants and course-level determinants.

Instructor-level determinants

Several key studies since the 1970s have outlined the instructor-level determinants that are key to being a "good" teacher. Characteristics students regard as meeting criteria for an "ideal" teacher map directly to attributes that closely match traditional criteria of teaching competency (Pozo-Munoz, Rebolloso-Pacheco, & Fernandez-Ramirez, 2000). Pozo-Munoz et al. 2000 identify the four main factors contributing to perceptions of good teaching as teaching competency, teaching qualities, teacher appearance, and directiveness. Analysis of the data shows that the most crucial attributes were related to teaching competency (which included "expertise", "informed", "clear", "able to motivate" among 16 other attributes) the least important related to directiveness (which included "prestigious", "attractive", "kind" (Pozo-Munoz et al., 2000) and other characteristics related to personality and personal appearance. Factors include clarity in teaching (Hativa, 1998; Ting, 2000); showing good management of student behaviour; excitement, and demonstrating interpersonal skills (Lowman & Mathie, 1993); being able to provide intellectual stimulation (Ting, 2000); showing a caring nature, and being systematic (Brown & Atkins, 1993); showing respect for students and being organised and having good presentations skills (Patrick & Smart, 1998).

Aigner and Thum 1986 found that 65% of the variation in the evaluation ratings could be explained by instructor-specific characteristics. The enthusiasm of the teacher, the level of interest stimulated, and the teachers' interaction with the students were among the most important instructor attributes. DeCanio 1986 found that communication skills and the level of organisation of the lecturer to be the most important characteristics. Boex 2000 used factor analysis to define six composite attributes of instructors: presentation skills, organisational skills and clarity of expression, how the instructor used grading and assignments, intellectual or scholarly capabilities, the ability to interact well with students, and the ability to motivate students. He noted that student evaluations were most likely to be influenced by the lecturer's organization, clarity, ability to motivate students, and the grading practices used by the lecturer. Mehdizadeh (1990) noted that expected grades, usefulness of supporting materials and the lecturer's help outside class to be critical to the evaluation results. In a study of the teaching of economics, the students evaluated their instructors' effectiveness on the bases of presentation, organisation, knowledge, accessibility, responsiveness and enthusiasm (Gokcekus, 2000).

Williams and Ceci's (1997) summary of previous findings indicated that, while ratings are reliable and do not change with student age, they are significantly influenced by instructor personal characteristics such as variations in voice patterns, warmth, supportiveness, dominance and confidence rather than attractiveness, gender or age. Other studies show that student evaluations are not unduly influenced by the instructor's personality and popularity or ability to entertain (Costin, Greenough, & Menges, 1971; Marsh & Ware, 1982; McKeachie, 1978). On the other hand Shevlin et al.'s (2000) recent study demonstrates that lecturer "charisma" is a significant underlying variable influencing student evaluation rankings, and they have argued that—as a consequence—evaluation surveys are not accurate measurements of teacher effectiveness. Radmacher and Martin (2001) found that teacher's extraversion was the only significant predictor of student evaluations after controlling for enrolment status, course grades and student ages. Felton, Mitchell and Stinson (2004) reach a similar conclusion in relation to 'sexy' rated professors. Students do appreciate instructors who are knowledgeable, warm, outgoing and enthusiastic (Murray, Rushton, & Paunonen, 1990). However, these same traits are likely to make the person a more effective teacher, so that students are stimulated to greater achievement and learning. If students feel they have learned they will give higher ratings. So the important factor is not how entertaining the instructor is. 'Neither the "stand up comic" with no content expertise nor the "cold fish" expert with only content expertise receives the highest ratings consistently' (Braskamp & Ory, 1994 p. 180).

Overall evaluations tended to be negatively correlated with age and with years of teaching experience (Wolfer & Johnson, 2003). These studies indicated a negative correlation between ratings of teachers and their teaching experience in that assistant professors were rated higher than full professors. However title, degree and position of an instructor: teaching assistants, visiting professors, tenure-track assistant professors or tenured professors, do not influence students' evaluations of them. Administrative experience, research publication and teachers' rank also have little effect on ratings (Ting, 2000).

The reputation of a subject instructor if known by students before enrolling in a subject, was shown to influence student ratings; more highly reputed instructors received higher ratings than did others (Griffin, 2000). In yet another study, economics students gave higher ratings on all evaluated items of teaching effectiveness to those instructors who spoke English as a first language than to those for whom English was not their native language (Bosshardt & Watts, 2001).

Student-level determinants

Given that student ratings may be made on the basis of features outside the lecturer's control a

number of studies have attempted to determine if student ratings do accurately measure teaching. The presence of biases in student ratings has been subject to much study.

One key source of potential bias is in relation to the influence of gender on teaching ratings however results are inconclusive. McKeachie (1979) and (Feldman, 1993) for example, found no significant effect for gender on student evaluations and Cashin's 1995 review of the literature supported this view. Other studies, though, have reported some effect for gender. Basow and Silberg (1987) surveyed the perceptions of more than 1000 students from the areas of humanities, social sciences, natural sciences and engineering of the effectiveness of 16 pairs of male and female instructors who were matched on the basis of rank, type of course taught and years of experience. They found that both male and female teachers were rated as effective, but whereas female students rated male and female teachers similarly, male students gave significantly less positive ratings to female teachers. Basow (2000) also requested students to describe their "best" and "worst" teachers, and found that, overwhelmingly, female students chose a female teacher as "best". When describing "worst" teachers, however, no similar disproportion was evident . Wolfer and Johnson (2003), in a study involving social work students, found that courses taught by female instructors were rated significantly higher than those taught by males. They suggested, however, that such a result might be explained by the fact that students tend to rate instructors of the same gender higher; approximately 90% of the participants in their study were female. In contrast, other studies have revealed that male instructors have been rated more highly than females. A review of research on college students' preconceptions of male and female college teachers revealed that in the majority of studies students' global evaluations of male and female teachers as professional were not different, though in a minority of studies male teachers received higher overall evaluations than did female teachers (Feldman, 1992). More recently, a study of 769 student evaluations of teaching in a Canadian university department of sociology and social studies found students rated male instructors more highly than female (Fries & McNinch, 2003).

Age of the student is another potential source of bias in student ratings. Centra (1993) found that the age of a student had no effect on student ratings. However, Worthington (2002) found that the age of the student did have some influence on the ratings given; students who were over the age of 30, and were also female, were more likely to assign a lower rating to the instructor.

Aptitude of students has been found to have some effect on the evaluations made of instructors. A study of economics subjects and teachers by Mason et al. (1995) revealed that better students were tougher in rating the quality of the subject, but more lenient in their rating of lecturers.

Positive but low correlations have been reported between student ratings and expected grade. That is, students expecting high grades in a subject tended to give higher ratings than did students expecting lower grades (Aigner & Thum, 1986; Mehdizadeh, 1990; Millea & Grimes, 2002) Worthington (2002), investigating the influence of student characteristics on the probability of particular rankings of one lecturer in a finance subject, made similar findings. Higher ratings were likely to be given by students who were expecting a higher grade and were from a non-English speaking background.

Grimes, Millea & Woodruff (2004) examined the relationship between student evaluations of their instructors and the innate and personal psychological construct of locus-of-control, a construct which identifies an individual's belief in their control over the environment. Those, for example, who believe they have little control over personal outcomes are categorised as 'externally oriented', whereas 'internally oriented' are those who accept responsibility for such control. Grimes et al. (2004) found that students identified as internally oriented are more likely to evaluate instructors highly than are externally oriented students who, believing that they had little personal control over

their grade, blame outside factors for their performances and so assign lower ratings to the instructor.

Subject-level determinants

Most studies on the influence of class size on student perceptions of quality of teaching have found little, if any, effect on the evaluation of teaching effectiveness (Marsh, 1987). Although there is a tendency for smaller classes to receive higher ratings it is a very weak inverse relationship (Cashin, 1995; Sixbury & Cashin, 1995). However, an Australian study conducted across four discipline groupings: humanities, sciences, social sciences and professional, found that larger classes were rated lower (Neumann, 2000). A class size effect has also been found by (Feldman, 1984; Liaw & Goh, 2002).

Ratings of perceived teaching quality do not appear to be related to the time of day or part of the year when the course is conducted (Liaw & Goh, 2002) or when the ratings are made (Abrami, Leventhal, & Perry, 1982; Cuseo, 2002; Seldin, 1993). The time during the term when ratings are collected also does not affect student ratings, and collections made at anytime during the second half of the semester yield similar ratings (Feldman, 1979). Nevertheless, Cuseo (2002) suggests that rating surveys administered immediately after a final examination might have less validity overall, due to the preoccupation, anxiety or fatigue of the students at that time.

Ratings by individual students of teachers do not change over time, with student post-subject experiences, or with the greater maturity of students. In cases where students have given retrospective ratings of subjects completed up to five years earlier, substantial agreement has been found between the retrospective ratings and those made at the time of the subject completion (Feldman, 1989; Feldman, 1989; Overall & Marsh, 1980). Such findings refute the argument that, because they are immature, students will come to appreciate subjects or instructors that were initially rated poorly, only with greater maturity.

Several studies have indicated the existence of a relationship between the nature of the discipline and student ratings. Feldman (1978) found Humanities and Arts-type subjects receive higher ratings than Social Science-type courses and that these, in turn, received higher ratings than Mathematics-type courses. Others have found similar results (Braskamp & Ory, 1994; Cashin, 1990; Centra, 1993; Marsh & Dunkin, 1992; Neumann, 2000; Sixbury & Cashin, 1995, 1995). Nevertheless, Cashin (1988), recommended caution in the interpretation of such findings, suggesting that, if instructors in fields requiring quantitative reasoning skills were rated lower because students were less competent in those skills, a bias existed in student ratings. Such a bias, Cashin argued, should be corrected. However, Marsh (1982), in a study designed to estimate the effects of normally confounded variables, concluded that the instructor is the primary determinant of the student rating, with an effect about five times as large as the effect of the subject.

Others have suggested that the type of instructional method employed can influence students' evaluation of the instructor. In a Canadian medical school study of three lecturers described as 'content-driven', 'context-driven' and 'pedagogy-driven' (p.99-100), Saroyan and Snell (1997) found that the student-centred context-driven and pedagogy-driven lectures were more highly rated than the teacher-centred content-driven lecture.

The issue of the influence of grades on student evaluations is an important one in the literature. Studies of a number of undergraduate courses found that grading leniency of instructors, together with expected workload in a subject, influenced the ratings of students (Greenwald & Gillmore, 1997). It has been suggested that some lecturers manipulate grades to obtain higher evaluation scores (Nelson and Lynch 1984; Zangenehzadeh, 1988, and Krautmann and Sander, 1999) Howard

and Maxwell (1980, 1982) however found little causal evidence of grading leniency and student satisfaction. Conversely, in Hong Kong (Ting, 2000), a teacher who gave higher grades was likely to receive lower ratings from students. There, it is suggested that as well as deeming such grading to be unfair, Chinese students rated teachers in ways consistent with traditional Chinese thinking, believing that good teachers 'should impose stringent standards on students' (Ting, 2000 p. 649).

In terms of the year level of students Mason et al. (1995) could attribute no significant difference to the year level of students in their ratings of teachers. Others however find that higher level subjects tend to receive higher ratings; first-year subjects, for example, were not rated as highly as those of later years while ratings for graduate subjects were higher than those for undergraduate subjects (Marsh & Bailey, 1993; Neumann, 2000). However the differences tend to be small (Neumann, 2000).

Students have been found to give lower ratings when they perceived the purpose of the evaluation process to be related to matters of staff tenure, but higher ratings when the purpose of the evaluation was seen to be the improvement of teaching in the future Worthington (2002).

The range of findings in relation to instructor, student and subject level influences on quality of teaching ratings by students prompted researchers in the current study to establish the experience in a department of economics at an Australian university. As has been demonstrated, the appropriate interpretation of existing evidence is problematic. This has led some academics to dismiss entirely any evidence gained from the surveys. Given the wide range of purposes for which this data is increasingly used within the department, the faculty and the university this seemed a dubious path to take and an analysis of our specific context was seen to be important. A detailed description of the study is reported elsewhere (Davies, Hirschberg, Lye, McDonald, & Johnston, 2005), however in this paper the key findings are reported and the implications for future practice are explored.

Overview of the current study

Student perceptions of the quality of teaching in the university have been undertaken for each subject, each semester since 1994. This study includes economics and econometric subjects in the analysis and uses an 8 year panel set data derived from the surveys conducted between semester 2 1995 and semester 2 2003.The data is derived from 3 sources: the average student rating scores as reported for all subjects taught in the Department of Economics; student records for all students who have taken an economics subject and additional subject-specific information as to the nature of the subject, its class size, time taught and other such information. Our aim was to establish how those factors, which are not under the control of the instructor, have an impact on the average student rating score attributed to their teaching efforts. We developed a statistical methodology, not discussed here, whereby we can condition these average scores in order to make comparisons across instructors.

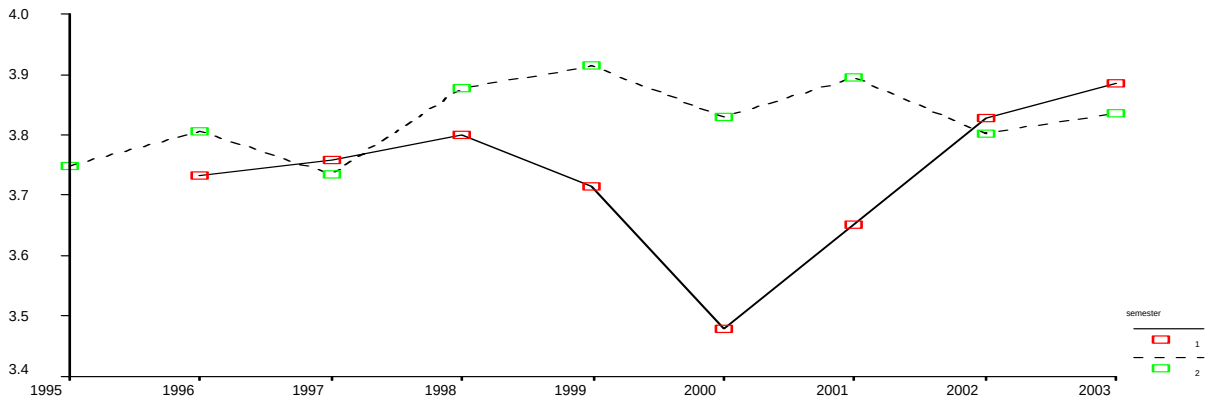
Of the 9 statements in the student rating surveys that students complete each semester for each subject and in which they are asked to rate their agreement five are highly correlated. While the wording of these statements changed over time the high degree of correlation indicates that they are eliciting the same information from students.

code	1995 to 1996	1997 to 2003
q1	The aims of the subject were made clear.	I had a clear idea of what was expected of me in this subject.
q2	This subject was well taught.	This subject was well taught.
q3	This subject was intellectually stimulating.	This subject was intellectually stimulating.
q5	In this subject, teaching staff showed an interest in the academic needs of students.	In this subject, teaching staff showed an interest in the academic needs of students.

code	1995 to 1996	1997 to 2003
q8	The explanations by the instructor of concepts and details were clear.	The explanations by the instructor were clear.

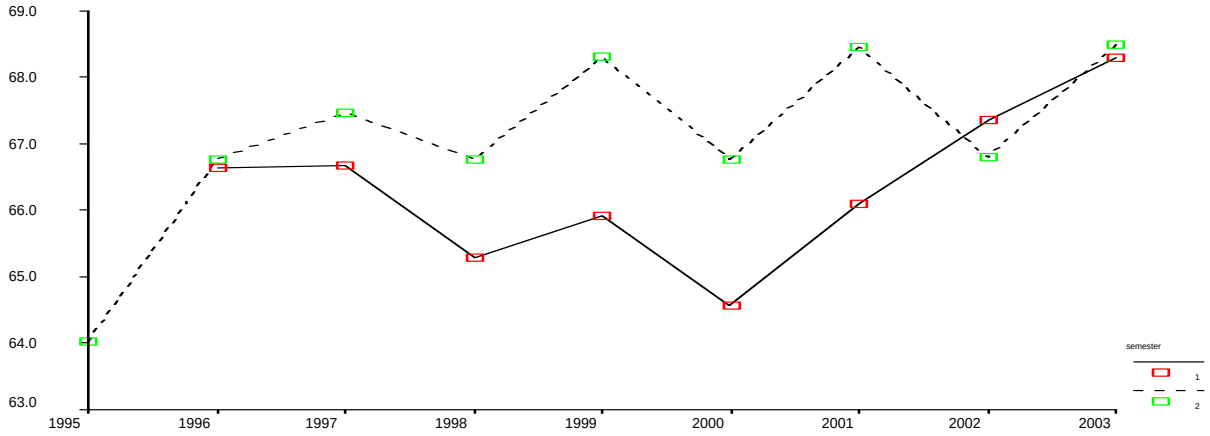
Of the five highly correlated questions, question 2 “This subject was well taught” is pivotal in the sense that this is the one most often referred to in confirmation and promotion decisions. Figure 1 shows the mean of this question between 1996 and 2003 indicates that which semester as subject is taught in can have an influence on questions 2. In most years to 2002 subjects taught in semester 2 score more on this question than subjects taught in semester 2. The change in 2002 may reflect the fact that many large and or compulsory subjects are now taught in both semesters. The difference in means between the two semesters would provide some support for this view.

Figure 1: Question 2 mean.



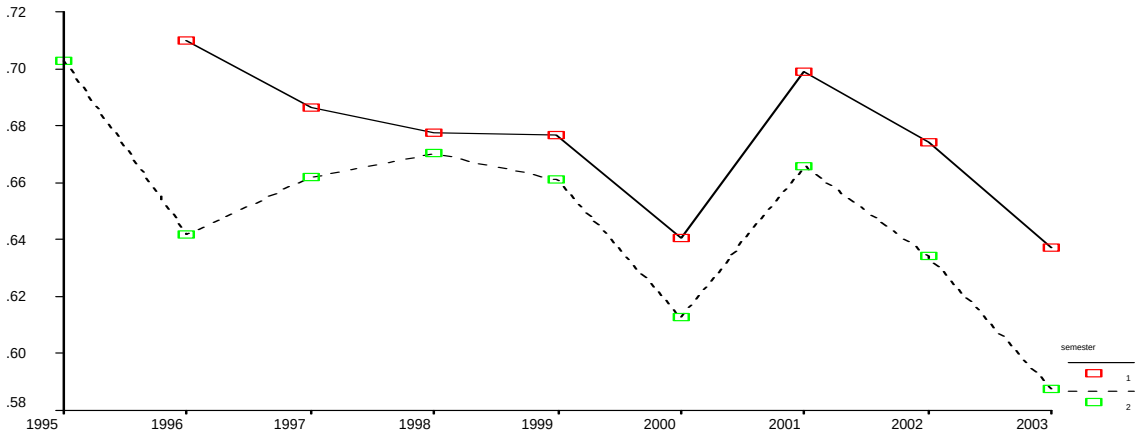
This semester difference is also reflected in the average mark over all student enrolments for each semester over the time period (Figure 2). Until 2002 the average marks across student enrolments in 2nd semester are higher than in first semester. The difference between semester average marks narrows markedly in 2002. Higher marks in semester 2 than in semester 1 may lend weight to the view that higher marks lead to higher Q of T scores

Figure 2: Average mark over all enrolments.



There has been a decline in the number of students responding to the surveys over the period 1996 to 2003 although this decline has been erratic. Figure 3 plots the proportion of QOT forms returned over all enrolments for each semester. The low level of response in 2003 is of concern and is likely to be related to a survey fatigue factor as well as a possibility that students are not well aware of the actions that are taken in response to survey results.

Figure 3: Proportion of forms returned over all enrolments



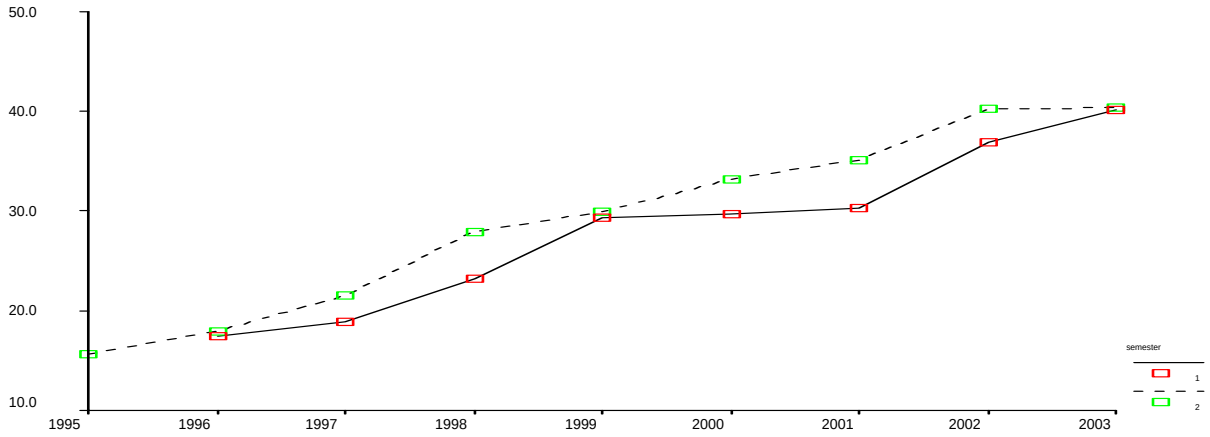
We sought to identify how much of the variation in student rating scores were due to the factors that were within the control of the lecturer and factors that are outside their control. Subject characteristics outside the control of the lecturer that were examined were class size, year level, and the nature of the subjects specifically whether it was quantitative or not. In relation to the student characteristics outside the control of the lecturer we looked at the proportion of students who were female and the proportion who belonged to a particular fee type ie international fee paying.

Results of an ordinary least squares regression analysis indicate that class size in first year subjects is not significant in determining student ratings of the quality of their teaching but larger class sizes in 3rd, 4th and graduate classes are negatively significant. In so far as a subject is quantitative we found that quantitative subjects scored marginally significantly positive result than non quantitative subjects. However when class size is accounted for larger sized quantitative classes (250 and above) are associated with lower student ratings than non quantitative subjects of the same size. Prior studies have found that mathematical subjects were more likely to be rated lower. (Feldman 1978, Braskamp and Ory 1994 Cashin 1990, Neumann 2000 etc) so the fact that we do not find this in our first and second year subjects is of interest.

In terms of student characteristics we found that the proportion of females in the subject was marginally negatively significant. This is at odds with McKeachie 1979, Feldman 1993 and Cashin 1995 who found no significant effect. However as there is wide variation in the literature in relation to the effects of gender and as our finding were marginal the results remain relatively inconclusive.

The proportion of international students has increased over the period (Figure 4) and it was therefore of interest to see if international students rated their lecturers differently than local students. We found that the proportion of students whose country of birth is Hong Kong or Singapore has a negative significant impact on student ratings whereas the proportion of students whose country of birth was the UK or Western Europe had a positive significant impact. Worthington (2002) on the other hand found that students from non-English speaking backgrounds were more likely to give higher ratings. However, for a number of countries from which a significant proportion of students originate (Malaysia, China, Indonesia, Indian sub-continent) there is no significant impact on QOT scores. This indicates that student learning cultures and expectations differ among students from Asian countries. More work on the impact of cultural backgrounds on student ratings needs to be completed before conclusions can be drawn in this regard.

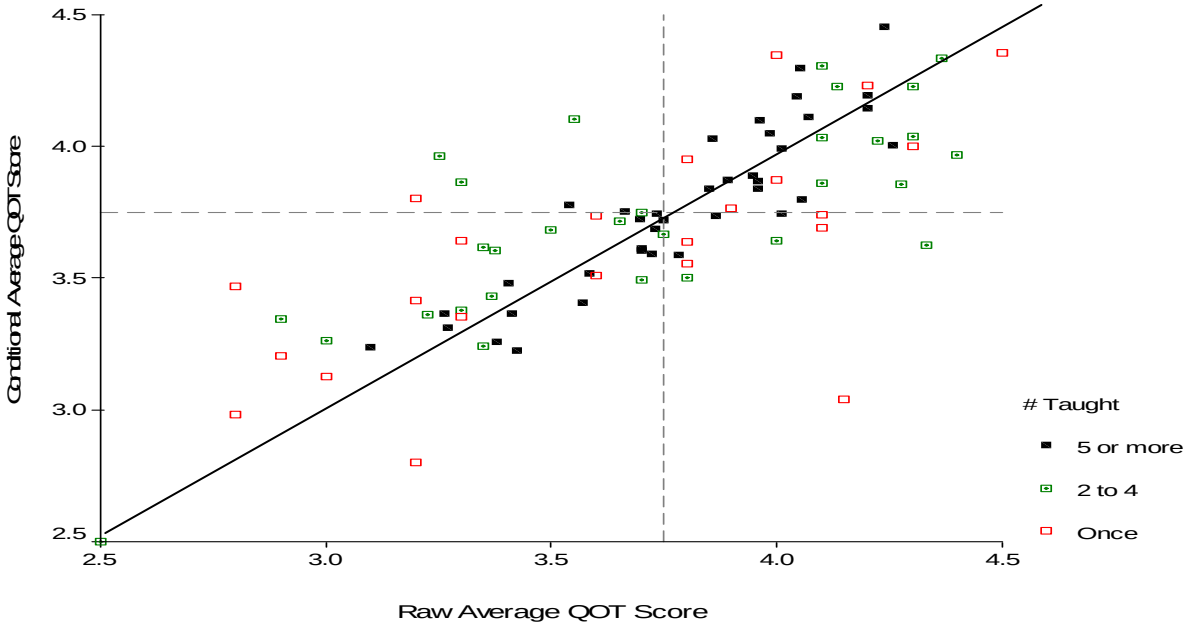
Figure4: % Foreign Fee over all enrolments



In relation to grade inflation our findings supported those of Millea and Grimes (2002 and Mehdizadeh (1990) that students expecting higher grades score their teachers higher. We also find that how well the other instructors of students in the class perform that year has an effect. Students seem to complete surveys in light of their experience with other instructors in the subjects they are enrolled in that year. Thus we find evidence that comparisons between the instructors take place.

These regression results can be used to adjust the mean for each instructor to provide a conditional mean where the subject and average student conditions are accounted for. The points on the 45° line indicate that class and student specific factors have no influence on the instructors' scores. In addition in the data set there are instructors with a wide range of teaching experience. To see if there are distinguishing characteristics as to who fell below the 45^{o} line we have highlighted those instructors who had only taught 1 subject in the data set, 2-4 subjects and more than 5. These are shown in Figure 5. We can see that if instructors had only taught 1 subject they were more likely to be further away from the 45^{o} line indicating that experience has a positive impact on student rating. This is to be expected. It would therefore seem sensible to look at an average of overall QOT scores rather than one particular value.

Figure 5: Raw Average vs. Conditional Average QOT Scores by Number of Courses taught



Conclusions

The data set that has been used in this study is large and has been collected over an extended time period. As a result some confidence may be placed in the reliability of the findings. In terms of those factors outside the control of the instructor that relates to subject level determinants we find that class size did not have a significant effect on student rating in first and second year subjects but did have a negative effect in later year classes. While other studies have indicated little effect of class size on rating s the difference between year levels in our study is worth noting. Other studies have found mathematical subjects tend to be rated lower but in our study we did not find this to be a factor except in quantitative subjects where enrolment exceeded 250. In terms of student characteristics that are outside the control of the instructor we found that the number of females in a subject to be marginally negatively significant. However this result is not sufficient to allow us to reach any meaningful conclusions about the effect of gender particularly in light of the wide range of results that have already be found in the existing literature. We did find that international students from some countries were more likely to rate their lecturers negatively. This aspect of student ratings could be a useful aspect of future research. The finding that students make comparisons between instructors and perceptions of good or bad teaching in other instructors influence the way in which students rate all instructors in that year is of interest in so far as it indicates that all instructors should be concerned that all reach a standard of teaching competency if all are to benefit.

The decline in the response rate is also of concern. Improved ways need to be found to give to students information on the results of the ratings and actions that have been taken so that they can appreciate the value of their feedback .

We believe that a case has been made for alerting individual staff to some of the conditional factors that should be taken into account when interpreting quality of teaching student rating results. Promotion and confirmation committees should also be made cognizant of the findings of this and similar studies.

Bibliography

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not know. *Journal of Educational Psychology, 82*(2), 219-231.

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

Access Economics (2005) *Review of Higher Education Outcome Performance Indicators* Report for Department of Education, Science and Training, April Australian Government Printers

Aigner, D. J., & Thum, F. D. (1986). On student evaluation of teaching ability. *Economic Education, 17*(4), 243-265.

Ballantyne, R., Bain, J. D., & Packer, J. (1999). Researching university teaching in Australia: Themes and issues in academics' reflections. *Studies in Higher Education., 24*(2), 237-257.

Basow, S. A. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles: A Journal of Research, 43*(5/6), 407-417.

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*(3), 308-314.

Bedggood, R. E., & Pollard, R. J. (1999). Uses and misuses of student opinion surveys in eight australian universities. *Australian Journal of Education, 43*(2), 129-156.

Boice, R. (n.d.). Countering common misbeliefs about student evaluations of teaching. from http://www.missouri.edu/~petwww/evaluation/student_eval/common_misbeliefs.htm

Bosshardt, W., & Watts, M. (2001). Comparing student and instructor evaluations of teaching. *Journal of Economic Education, 32*(1), 3-17.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.

Brown, G., & Atkins, M. (1993). *Effective teaching in higher education*. London: Routledge.

Byrne, M., & Flood, B. (2003). Assessing the teaching quality of accounting programs: An evaluation of the course experience questionnaire. *Assessment & Evaluation in Higher Education, 28*(2), 135-145.

Casey, R. J., Gentile, P., & Bigger, S. (1997). Teaching appraisal in higher education: An australian perspective. *Higher Education, 34*(4), 459-482.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research. Idea paper no. 20*.Kansas.

Cashin, W. E. (1990). Students do rate different academic fields differently. In M. Theall & J. Franklin (Eds.), *Students ratings of instruction: Issues for improving practice: New directions for teaching and learning* (pp. 113-121). San Francisco: Jossey-Bass.

Cashin, W. E. (1995). Student ratings of teaching: The research revisited. *IDEA Paper No. 32, 32*(Sept).

Cashin, W. E., & Downey, R. G. (1992). Using global student ratings items for summative evaluation. *Journal of Educational Psychology, 84*(4), 563-572.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*.San Francisco: Jossey-Bass.

Costin, F., Greenough, W. H., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 8*(511-35).

Cuseo, J. (2002). Evaluating new-student seminars and other first-year courses via course-evaluation surveys: Research-based recommendations regarding instrument construction and administration, data analysis, data summary, and reporting results. *http://www.brevard.edu/fyc/fya/CuseoLink.htm*.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208.

Department of Employment, Science and Training (2004). *Our Universities: Backing Australia's future - The Higher Education Reforms,* Canberra, Australian Government printer.

Davies, W. M., Hirschberg, J., Lye, J., McDonald, I., & Johnston, C. (2005). Systemic influences on teaching evaluations: The case for caution.

Dwinell, P. L., & Higbee, J. L. (1993). Students' perceptions of the value of teaching evaluations. *Perceptual and Motor Skills, 76*, 995-1000.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education, 9*, 199-242.

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education, 10*(2), 149-172.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education, 21*(1), 45-116.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*(6), 583-645.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students colleagues, administrators and external (neutral) observers. *Research in Higher Education, 30*(2), 137-.

Feldman, K. A. (1992). College students' views of male and female college teachers: Part 1 - evidence from the social laboratory and experiments. *Research in Higher Education, 33*, 317-375.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part ii--evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*(2), 151-211.

Felton, J. Mitchell, J & Stinson, M. (2004).Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness , *Assessment & Evaluation in Higher Education*, 29 (1) Feb 92-107

Fries, C. J., & McNinch, R. J. (2003). Signed versus unsigned student evaluations of teaching: A comparison. *Teaching Sociology, 31*(3), 333-344.

Gokcekus, O. (2000). How do university students value economics courses? A hedonic approach. *Applied Economics Letters, 7*, 493-496.

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

Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*(4), 743-751.

Griffin, B. W. (2000). Instructor reputation and student educational psychology. from http://coe.gasou.edu/foundations/bwgriffin/research/instructor_reputation_2000.htm
retrieved 8/10/2003

Grimes, P. W., Millea, M. J., & Woodruff, T. W. (2004). Grades--who's to blame? Student evaluation of teaching and locus of control. *Journal of Economic Education, 35*(2), 129-147.

Hativa, N. (1998). Lack of clarity in university teaching: A case study. *Higher Education, 36*(3), 353-381.

Hativa, N., & Raviv, A. (1993). Using a single score for summative evaluation by students. *Research in Higher Education, 34*(5), 625-646.

Haynes, B. (2002). *The use of student evaluations as an indicator of teaching quality in higher education.* Paper presented at the Conference of Australian Association for Research in Education, Brisbane.

Hepworth, D., & Oviatt, B. E. (1985). Using student course evaluations: Findings, issues, and recommendations. *Journal of Social Work Education, 21*(3), 105-112.

Howard,G.S.& Maxwell S.E. (1982) Do grades contaminate student evaluations of instructions? Research in higher education ,21(2), 187-206

Husbands, C. T. (1996). Variations in students' evaluations of teachers' lecturing and small-group teaching: A study at the London school of economics and political science. *Studies in Higher Education, 21*(2), 187-206.

Husbands, C. T. (1997). Variations in student evaluations of teachers' lecturing in different courses on which they lecture: A study at the London School of Economics and political science. *Higher Education, 33*(1), 51-70.

Husbands, C. T., & Fosh, P. (1993). Students' evaluation of teaching in higher education: Experiences from four European countries and some implications of the practice [1]. *Assessment & Evaluation in Higher Education, 18*(2), 95-114.

Kolitch, E., & Dean, A. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about 'good' teaching. *Studies in Higher Education, 24*(1), 27-42.

Liaw, S. H., & Goh, K. L. (2002). Biases in student evaluation of teaching: The case of faculty of economics & administration university of Malaya. *FEA Working Paper, No. 2002-3*.

Low, G. (1999). What respondents do with questionnaires: Accounting for incongruity and fluidity. *Applied Linguistics, 20*(4), 503-533.

Lowman, J., & Mathie, V. A. (1993). What should graduate teaching assistants know about teaching? *Teaching of Psychology, 20*(2), 84-88.

Marsh, H. W. (1982). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement, 6*, 47-59.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253-388.

Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness. *Journal of Higher Education, 64*(1), 1-18.

Marsh, H. W., & Dunkin, M. (1992). Students' evaluations of university teaching: A multi-dimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook on theory and research* (pp. 143-234). New York: Agathon Press.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7*, 300-314.

Marsh, H. W., & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30*(1), 217-251.

Marsh, H. W., & Ware, J. E., Jr. (1982). Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. *Journal of Educational Psychology, 74*(1), 126-134.

McCormick, C. (1996). *A focus group approach to undergraduate psychology program assessment*.Georgia.

McKeachie, W. J. (1978). *Teaching tips: A guidebook for the beginning college teacher*.Lexington, MA: Heath.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe: Bulletin of the AAUP, 65*(6), 384-397.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225.

McKeachie, W. J., & Lin, Y. G. (1979). A note on validity of student ratings of teaching. *Educational Research Quarterly, 4*(3), 45-47.

Mehdizadeh, M. (1990). Loglinear models and student course evaluations. *Journal of Economic Education, 21*(1), 7-21.

Millea, M., & Grimes, P. W. (2002). Grade expectations and student evaluation of teaching. *College Student Journal, 36*(4), 582-590.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology, 82*(2), 250-261.

Nelson, J.P., & Lynch , K.A., (1984) Grade inflation, Real income, Simultaneity and teaching Evaluations . *Journal of Economic Education*, 15(1), 21-37

Neumann, R. (2000). Communicating student evaluation of teaching results: Rating interpretation guides (rigs). *Assessment & Evaluation in Higher Education, 25*(2), 121-134.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72*(3), 321-325.

Patrick, J., & Smart, R. M. (1998). An empirical investigation of teacher effectiveness: The emergence of three critical factors. *Assessment & Evaluation in Higher Education, 23*(2), 165-178.

Pozo-Munoz, C., Rebolloso-Pacheco, E., & Fernandez-Ramirez, B. (2000). The 'ideal teacher'. Implications for student evaluation of teacher effectiveness. *Assessment & Evaluation in Higher Education, 25*(3), 253-263.

Radmacher, S. A., & Martin, D. J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *The Journal of Psychology, 135*(3), 259-268.

Saroyan, A., & Snell, L. (1997). Variations in lecturing styles. *Higher Education, 33*(1), 85-104.

Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education, 39*(3), 40.

Sheehan, D. S. (1975). On the invalidity of student ratings for administrative personnel decisions. *Journal of Higher Education, 46*(6), 687-700.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397-405.

Simpson, P. M., & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education, 22*(3), 199-213.

Sixbury, G. R., & Cashin, W. E. (1995). Comparative data by academic field. In *Idea technical report, vol. 10*:Center for Faculty Evaluation and Development, Manhattan, Kansas: Kansas State University.

Sixbury, G. R., & Cashin, W. E. (1995). Description of database for the idea diagnostic form. In *Idea paper, vol. 9*:Center for Faculty Evaluation and Development, Manhattan, Kansas: Kansas State University.

Smith, G. S. (2004). Assessment strategies: What is being measured in student course evaluations? *Accounting Education, 13*(1), 3-28.

Solas, J. (1990). Effective teaching as construed by social work students. *Journal of Social Work Education, 26*, 145-154.

Ting, K. (2000). A multi-level perspective on student ratings of instruction. *Research in Higher Education, 41*(5), 637-661.

Wachtel, H. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191-211.

Wagner, Z. M. (1999). Using student journals for course evaluation. *Assessment & Evaluation in Higher Education, 24*(3), 261-272.

Williams, W. M., & Ceci, S. J. (1997). "How'm I doing?" problems with student ratings of instructors and courses. *Change, 29*(5), 12-23.

Wilson, R. (1998). New research casts doubt on value of student evaluations of professors. *The Chronicle of Higher Education*(16 January, 1998), A12.

Wolfer, T. A., & Johnson, M. M. (2003). Re-evaluating student evaluation of teaching: The teaching evaluation form. *Journal of Social Work Education, 39*(1), 111-121.

Worthington, A. C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education. *Assessment & Evaluation in Higher Education, 27*(1), 49-64.
Zangenehzadeh, H. (1988) Grade inflation: A way out. *Journal of Economic Education* 19(3), 217-230