The International Journal for Academic Development

ISSN: 1360-144X (Print) 1470-1324 (Online) Journal homepage: http://www.tandfonline.com/loi/rija20

Does evaluation of teaching lead to improvement of teaching?

Harry G. Murray

To cite this article: Harry G. Murray (1997) Does evaluation of teaching lead to improvement of teaching?, The International Journal for Academic Development, 2:1, 8-23, DOI: 10.1080/1360144970020102

To link to this article: https://doi.org/10.1080/1360144970020102

Does evaluation of teaching lead to improvement of teaching?

Harry G. Murray, *Department of Psychology, University of Western Ontario, Canada*

ABSTRACT

Given the widespread use of student evaluation of teaching in North American colleges and universities, it is reasonable to ask whether student evaluation has accomplished one of its major intended outcomes, namely improvement of instructional quality. A review of research evidence from three independent sources (faculty surveys, field experiments and longitudinal comparisons) suggests that student evaluation does in fact contribute significantly to improvement of certain aspects of university teaching, particularly if evaluation is supplemented by expert consultation. Furthermore, there is no clear evidence that student evaluation has led to undesirable instructional side-effects, such as grade inflation and entrenchment of traditional teaching methods.

Introduction

Student evaluation of teaching represents one of the most important and most controversial developments in higher education in the past 50 years. Beginning in the late 1960s and early 1970s, colleges and universities in the United States and Canada began to use formal student ratings of teaching both as feedback to instructors and as input for administrative decisions on faculty salary, retention, tenure and promotion. More recently, this practice has become more common in other parts of the world, such as Australia, Britain, Nigeria, Thailand, Switzerland, Belgium, Hong Kong, Israel and New Zealand (Miller, 1988).

Most student evaluation forms in current use assess teacher and course characteristics such as *clarity of explanation, enthusiasm for subject* matter, encouragement of student participation, breadth of coverage, and quality of feedback, that are assumed to be:

- observable by students;
- under the control of the instructor; and
- correlated with student learning.

Although many faculty members continue to be vehemently opposed to the idea that students should be involved in evaluation of teaching, the general trend has been for such evaluation to be adopted more and more widely, to the extent that nearly 100% of higher education institutions in North America now make use of some form of student evaluation of teaching. In some cases, student evaluation is supplemented by other sources of data on teaching, such as input from colleagues, but in many institutions student ratings provide the sole documentation for quality of teaching.

One reason that student evaluations have increased in popularity and acceptability is that research evidence from over 1,500 published studies indicates student ratings can provide reliable and valid evidence of teaching effectiveness. Although findings vary somewhat from study to study, the weight of evidence suggests that student ratings of a given instructor:

- show high reliability or stability across items, groups of raters, time periods and courses;
- are affected to only a minor extent by extraneous factors such as class size and severity of grading;
- correlate significantly with comparable ratings made by colleagues, alumni and trained classroom observers; and
- most important of all, are positively related to more objective measures of teaching

effectiveness, such as student performance on a common, objectively scored final examination in a multi-section course (Cohen, 1981; Marsh & Dunkin, 1992; Murray, 1980).

Although most of the published evidence on the reliability and validity of student instructional ratings was conducted in North American universities, research conducted in other countries has yielded results generally very similar to those outlined above. For example, Watkins (1994) reported that student evaluations of teaching obtained in six culturally different countries (India, Nepal, Nigeria, Philippines, Hong Kong and New Zealand) were similar in terms of internal consistency reliability, convergent and discriminant validity, and factors that differentiated between good and poor teachers. Similarly, Baker (1986) found that Palestinian student ratings of teachers were unaffected by potential biasing factors such as expected grade and gender; and Prosser and Trigwell (1990) showed that Australian university students taught by highly rated teachers tended to use deep rather than surface study strategies.

The present paper is concerned not with the reliability and validity of student instructional ratings but with the equally important but relatively ignored question of whether student evaluation of teaching has contributed significantly to improvement of teaching. Given that teaching improvement is typically one of the main justifications given for the introduction of student ratings, it would be interesting to know whether such evaluations have in fact led to improvement in teaching quality. Messick (1989) points out that the quality of a measurement procedure is determined in part by the accuracy of information it provides and in part by the impact it has on the performance of those who are measured. Vogt and Lasher (1973) suggest that 'the ultimate product of student evaluation ought to be improved instruction.'

Certainly there are logical reasons for expecting that student evaluation *should* contribute to improvement of teaching. First, student ratings provide diagnostic feedback to teachers, and feedback is usually found to contribute to improved performance. Second, student evaluation can provide the impetus to seek expert consultation or participate in faculty development programmes aimed at improvement of teaching. Third, the use of student evaluations in faculty salary, tenure, and promotion decisions provides a tangible incentive to put time and effort into improvement of performance. Finally, the use of student evaluations in faculty personnel decisions (particularly those relating to hiring, retention and tenure) provides a selection mechanism whereby more effective teachers are more likely to be recruited and retained by an institution. Although these reasons are intuitively appealing, they do not constitute a convincing empirical demonstration that student evaluation of teaching does in fact lead to improved teaching.

The remainder of this paper reviews research evidence from three independent sources on the question of whether or not student evaluation of teaching has contributed to improved teaching, namely:

- surveys of faculty opinion;
- field experiments on the effects of student feedback; and
- longitudinal analyses of faculty teaching performance after the introduction of student evaluation.

Following this review, I conclude with an examination of possible negative side-effects of student evaluation of teaching, including grade inflation and entrenchment of traditional, teacher-centred methods of teaching.

Faculty Surveys

One method of assessing whether student evaluation has improved teaching is to survey the opinions of faculty members as to the formative value of feedback from students. Table 1 summarizes the results of eight published surveys of faculty opinion that asked one or both of the following questions: 'Do student ratings provide useful feedback for improvement of teaching?' and 'Have student ratings led to improved teaching?' Although results varied somewhat from study to study, the overall trend was for faculty respondents to agree that student ratings have had a positive impact on quality of teaching.

In the largest survey to date, Outcalt (1980) reported that 67% of 4,468 respondents at various campuses of the University of California said that student ratings had contributed to improvement of teaching. Similarly, 54% of 666 faculty respondents at the University of Western Ontario stated that

Table 1 *Surveys of faculty opinion on formative impact of student instructional ratings*

Survey	Number of respondents	Percent Agreement: Do student ratings provide useful feedback?	Percent Agreement: Have student ratings led to improved teaching?
Outcalt (1980) California, USA	4468	77	67
McCready (1981) Wilfrid Laurier, Canada	25	76	80
Gross & Small (1979) George Mason, USA	163		84
Owens (1977) Kansas State, USA	263		88
Jacobs (1987) Indiana, USA	96		70
Menges (1980) Northwestern, USA	193	73	
Murray et al. (1982) Western Ontario, Canada	666	54 (global ratings) 65 (prose comments) 78 (specific ratings)	
Ory & Braskamp (1981) Illinois, USA	25 22	54* (rating scales) 63* (prose comments)	

* Estimated from mean ratings on 5-point scale

student ratings of general teacher characteristics provided useful feedback, whereas 65% said that prose comments from students were useful as feedback, and 78% reported that student ratings of specific teaching behaviours were valuable for feedback purposes. Across all surveys reviewed in Table 1, and with differential weighting of sample size, 73.4% of respondents said that student evaluations provided useful feedback, and 68.8% said that student evaluations have led to improved teaching.

In a study not listed in Table 1 (because it included neither of the two questions identified above) Ryan, Anderson and Birchler (1980) asked instructors at the University of Wisconsin-Lacrosse to indicate whether student ratings had caused them to change their frequency of use of various instructional methods and practices. Instructors reported increased use of several practices that would normally be viewed as 'good teaching', such as explicit definition of objectives, availability for consultation, opportunity for classroom discussion and prompt return of exams and papers. Unfortunately, instructors also reported increased use of several undesirable teaching practices, such as watering down of course content, grade inflation, and decreased exam difficulty. In general, faculty members at Lacrosse felt that student ratings had not improved quality of teaching.

In summary, it appears that a clear majority of faculty members believe that student evaluation of teaching provides useful feedback and has led to improvement of teaching. The one exception to this generalization is the Ryan et al. (1980) study, but even it can be interpreted as a mixture of positive and negative results. It goes without saying, of course, that faculty opinion surveys have inherent limitations as a source of evidence on the impact of student instructional ratings. For one

thing, faculty members who voluntarily participate in survey research on student evaluation of teaching may be individuals who have inordinately positive (or negative) attitudes towards this topic. A second problem is that some faculty respondents may give 'socially desirable' answers rather than 'true' answers to survey questions, thus inflating the level of support found for student evaluation of teaching. A third possibility is that faculty endorsement of student ratings may represent some form of rationalization or dissonance reduction in response to completion of what for many faculty would be a rather unpleasant task!

Field Experiments

A second way of assessing the formative impact of student evaluation of teaching is to carry out a field experiment similar to that depicted in the top half of Figure 1, in which randomly assigned experimental teachers receive feedback concerning mid-course student evaluation of teaching, whereas control teachers are evaluated at mid-term but given no feedback. The two groups are then compared on end-of-course student ratings, with the expectation that experimental teachers will show higher ratings as a result of the beneficial effects of feedback. Since field experiments involve external (ie student) assessment of end-of-term teaching performance, they rule out errors and biases in self-report data that are problematic in surveys of faculty opinion.

In a variation on the basic feedback design shown in Figure 1, McKeachie et al. (1980) compared groups of teachers who, at mid-semester, received either no student feedback, a standard computer printout of student ratings, or a computer printout supplemented by individual consultation with a faculty development specialist who interpreted the printout, provided motivational support, and offered specific suggestions for improvement. The three groups differed significantly in end-of-semester student ratings, with the feedback-plus-consultation group showing the highest ratings, followed by the feedback-only group and the control group. In other words, the results indicated that student feedback alone led to modest improvement in perceived quality of teaching, whereas student feedback supplemented by expert consultation produced much larger gains in teaching.

Cohen (1980) and Menges and Brinko (1986) conducted meta-analytic reviews of student feedback field experiments and reached conclusions very similar to those of the McKeachie et al. study. The bottom half of Figure 1 summarizes the key findings of Cohen's review. It may be noted that 13 experiments compared feedback alone versus no feedback, and on average these experiments found a small but significant increment of 0.10 points on a 5-point scale (3.70 versus 3.80) in end-of-term ratings due to mid-term feedback. On the other hand, nine experiments compared student feedback plus expert consultation versus no feedback, and found an average increment in final ratings of approximately 0.33 points (3.70 versus 4.03). The gain in teacher ratings due to feedback plus consultation corresponds to approximately two-thirds of a standard deviation or 24 percentile points. Thus an instuctor starting at the 50th percentile in student ratings would be expected to improve to the 74th percentile as a result of mid-term student feedback plus expert consultation. Gains of this magnitude obviously cannot be dismissed as trivial. Moreover, the finding that student feedback plus expert consultation leads to greater improvement in teaching than student feedback alone provides empirical support for the contribution of faculty development offices and programmes to enhance college and university teaching. It appears that student evaluation and faculty development play complementary and synergistic roles in teaching improvement.

Like all forms of research, student feedback field experiments have weaknesses and limitations. One limitation is that most of the field experiments reviewed by Cohen (1980) were conducted with samples of experienced teachers who had already received student feedback in previous courses, so that experimental feedback effects were superimposed on prior feedback. It should be noted, however, that Centra (1973) reported a significant (but delayed) student feedback effect in a sample of teachers, who had no prior experience with student ratings of teaching, at five liberal arts colleges. A second limitation of field experiments is that they have tended to rely solely on end-of-course student ratings as a measure of teaching performance. However, Overall and Marsh (1979) found significant feedback effects with two alternative measures of teaching effectiveness, namely student

Research Design

		MID-TERM STUDENT FEEDBACK	END-OF-TERM STUDENT RATINGS
RANDOM ASSIGNMENT	CONTROL GROUP	NO	YES
	EXPERIMENTAL GROUP	YES	YES

Results

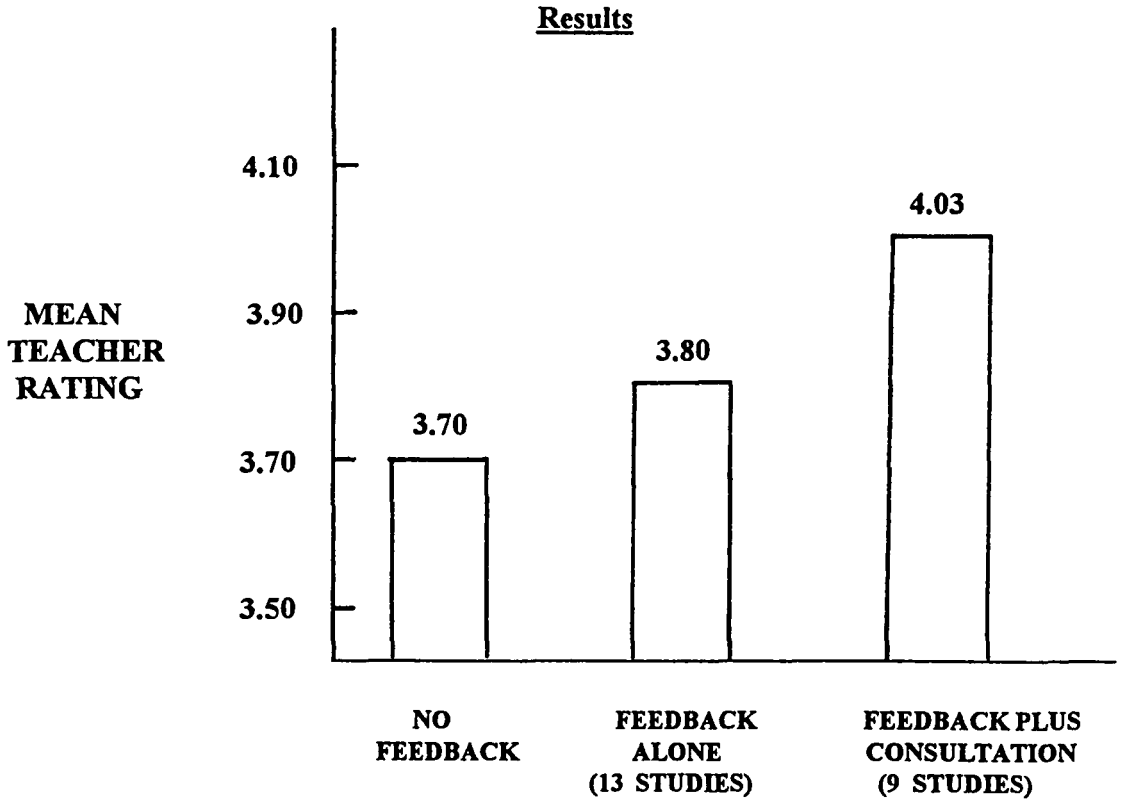

Figure 1 *Field experiments on effectiveness of student feedback* *Source: Cohen (1980)*

examination performance and subsequent course enrolment. A third limitation of field experiments is that improvement in teaching is usually demonstrated on a short-term basis only, with most experiments conducted over a time frame of at most two or three months. Stevens and Aleamoni (1985), however, reported effects of student feedback plus expert consultation that persisted for as long as ten years.

One possible reason for the limited impact of student feedback without expert consultation in field experiments is that the feedback provided was too vague and nonspecific to be useful for improvement (Murray, 1987a). Low student ratings on an item such as 'explains clearly' or 'has good rapport' tell the instructor that something is wrong, but provide no clear indication as to exactly what is wrong or specifically what needs to be changed to bring about improvement.

To remedy this situation, Murray and Smith (1989) carried out a field experiment in which graduate teaching assistants in English, geography, and psychology received mid-term student feedback on the frequency with which they exhibited more specific or fine-grained teaching behaviours such as 'signals the transition from one topic to the next' and 'addressses individual students by name'. The difference in mean end-of-term student ratings between experimental and control teachers was 0.38 points on a 5-point scale (3.73 versus 3.35), or three-quarters of a standard deviation. This is a much larger difference than that reported in field experiments using nonspecific feedback without consultation, and in fact is approximately equal to the average difference of 0.33 points found in experiments where global feedback was supplemented by expert consultation.

In summary, field experiments provide evidence that student feedback alone leads to modest improvement in faculty teaching performance, whereas student feedback supplemented either by expert consultation or by clarification of specific teaching behaviours leads to more substantial gains in quality of teaching. As noted above, the critical role of expert consultation in moderating the impact of student feedback suggests that faculty development programmes make a strong contribution to improvement of teaching. In response to such evidence, a number of writers have proposed systematic models for combining evaluation and consultation. For example, Wilson (1986) described a consultation procedure in which a teacher seeking to improve a particular dimension of teaching (eg organization, student participation) is provided with specific tips from outstanding, award-winning teachers who have received particularly high student ratings on that same dimension. In addition, the consultant conducts a complete analysis of the client's teaching evaluation data, sends a one-page written description of each of the relevant teaching tips, and phones the client periodically to check on progress to date.

Marsh and Roche (1993) conducted a field experiment assessing the impact of Wilson's evaluation/consultation model at the University of Western Sydney, Australia, a newly established institution that had had no prior experience with student evaluation of teaching. Randomly assigned groups of teachers received either mid-semester student feedback plus consultation, end-of-semester feedback plus consultation, or no intervention (control group); then all groups were compared on student ratings at the end of the *next* semester. It was found that all three groups showed improvement in student ratings across the three observation periods (mid-semester 1, end-semester 1, end-semester 2), but improvement was greater for targeted teacher rating items than for non-targeted items and was significant only for the end-of-semester group. Marsh and Roche concluded that student evaluation coupled with Wilson's (1986) model of consultation is an effective means of improving teaching.

Longitudinal comparisons

A third way of assessing whether evaluation of teaching leads to improvement of teaching is by comparing mean student rating scores longitudinally over a period of several years following the introduction of student evaluation in a particular academic unit (department or faculty). If student evaluation contributes to improvement of teaching, this improvement should be reflected in a gradual increase across years in the average teacher rating score for participating faculty members. Ideally, a valid test of this hypothesis requires that the following conditions be met:

- mean ratings are compared across a minimum of ten years or ten semesters;

- tracking of mean ratings across years begins in the same year that student evaluation was first introduced;
- the same student rating form is used throughout the study; and
- all faculty and all courses undergo student evaluation in all years.

Published research on longitudinal trends in student ratings of teaching has yielded mixed results. Of 14 studies located by the author, eight reported significant longitudinal improvement and six reported no significant change in student ratings over time. However, as outlined below, most studies conducted to date have failed to fulfil the four methodological conditions identified above. For example, Gray and Brandenburg (1985) found significant longitudinal improvement in mean student ratings of teaching in a sample of 304 faculty members from various academic disciplines at the University of Illinois, but ratings were tracked over only four consecutive semesters, and the study did not begin in the semester where student evaluation was introduced. Similarly, Pigott and Rosehart (1983) reported significant longitudinal improvement in student ratings following the introduction of mandatory teaching evaluation in six professional schools at Lakehead University, Canada, but ratings were tracked for only four successive semesters. Vogt and Lasher (1973) found no significant improvement in mean student ratings for a group of 50 instructors in the College of Business Administration at Bowling Green State University, USA, despite the fact that student evaluation was accompanied by a peer consultantship programme and student evaluation results were used on a mandatory basis in faculty promotion and tenure decisions. As in the Pigott and Rosehart study, longitudinal tracking was introduced concurrently with the advent of student evaluation, but mean ratings were compared across only eight academic quarters between 1969 and 1972.

In a large-scale study that fulfilled all of the four methodological conditions named above, Marsh and Hocevar (1991) found no evidence of longitudinal improvement in teaching following the introduction of student ratings. The sample of teachers in the Marsh and Hocevar study consisted of 195 faculty members from 31 different departments at the University of Southern California, each of whom had been evaluated in each of at least ten different years over a 13-year period from 1976 to 1988. All instructors were evaluated by the same evaluation form, namely the Students' Evaluations of Educational Quality (SEEQ) instrument, which was introduced in 1976 for mandatory use in promotion and tenure decisions. Ratings of a given instructor on each of the 11 SEEQ dimensions were averaged across all courses taught in the same year, and trends across years were assessed by multiple regression procedures. It was found that there was virtually no change in mean student ratings for the sample as a whole across the 13-year observation period. The correlation between year and rating was significant (but in a negative direction) for only two of eleven SEEQ dimensions, and year accounted for less than 1% of variance in student ratings. Follow-up analyses ruled out the possibility that possible improvement in faculty teaching was counterbalanced by a longitudinal increase in the standards used by students in evaluating teaching effectiveness. Thus, despite the use of a large sample and powerful design, the Marsh and Hocevar study provided no evidence that mean student ratings improve over time following the introduction of student evaluation of teaching.

Results very different from those of Marsh and Hocevar were reported by the present author (Murray, 1987b; Murray, Jelley, & Renaud, 1996) in a large-scale longitudinal study conducted in the Department of Psychology, University of Western Ontario. Like the Marsh and Hocevar study, the University of Western Ontario study fulfilled all of the four methodological conditions identified above. Student evaluation results were available for all courses taught in each of 26 consecutive academic years (1970 to 1995) by 40 to 50 full-time faculty members in the Department of Psychology. Furthermore, the same ten-item teaching evaluation form, which focuses on classroom presentational skills, such as clarity of explanation and use of examples, was used continuously throughout this 26-year period. Evaluation was done annually for all teachers and all courses, with results of evaluation used on a mandatory basis in promotion and tenure decisions. Figure 2 shows the average student rating of teaching (on a 5-point rating scale) for all faculty members in the Department of Psychology in each of 26 consecutive academic years from 1969–70 to 1994–95.

Student rating data for each faculty member were averaged across all items of the teacher

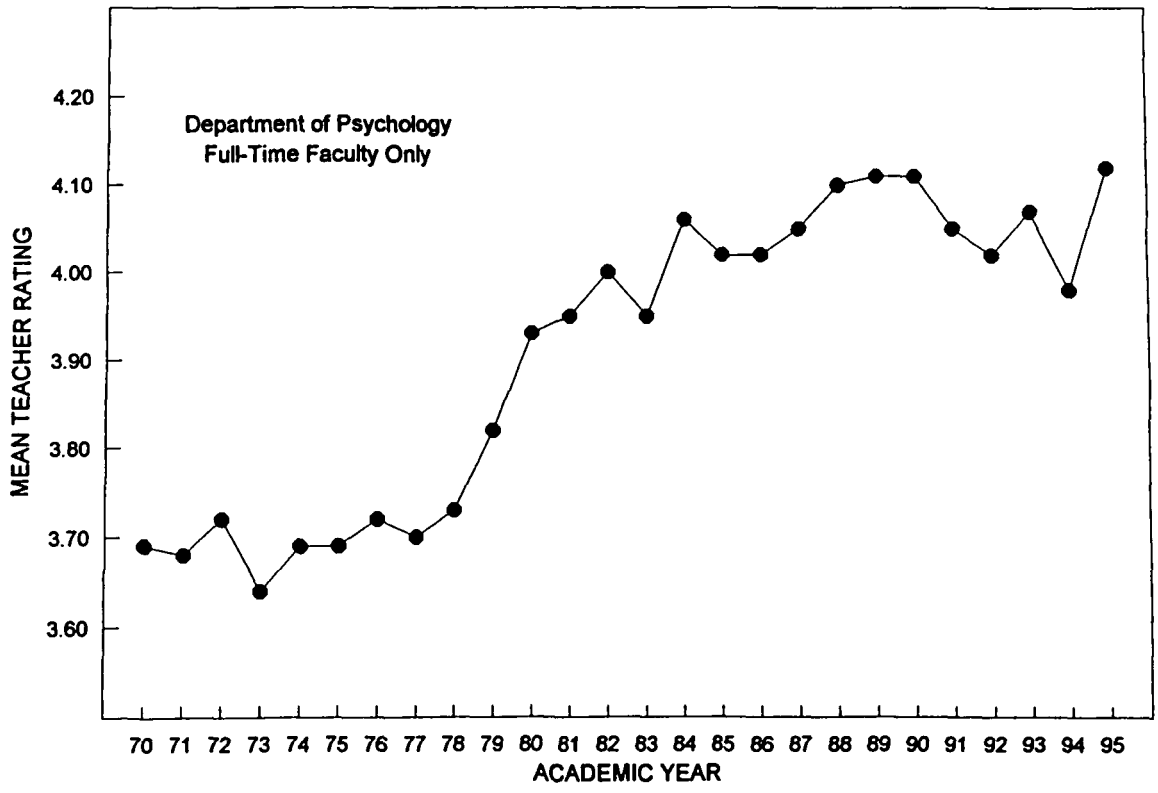

Figure 2 *Mean teacher rating scores for Department of Psychology as a whole for academic years 1969–70 to 1994–95.*

evaluation form and across all courses taught in a given year. It may be noted that the departmental average teacher rating increased from approximately 3.70 in the early 1970s to approximately 4.05 in the early 1990s, which corresponds to a gain of slightly more than 1.0 standard deviation units. A regression line fitted to the data points in Figure 2 (not shown) was found to deviate significantly from zero, and the correlation coefficient between year and department mean rating was 0.91. It may also be noted in Figure 2 that little or no improvement took place in the department mean rating over relatively long stretches of time, such as from 1970 to 1978 and from 1984 to 1995. The reason for this anomaly is not clear, but it does suggest that failure to find longitudinal improvement in previous studies that tracked mean student ratings over periods of only six to eight semesters (three to four years) may have been due to the use of too short an observation period.

It is also worth noting that Marsh & Hocevar (1991) used the individual teacher as the unit of analysis, whereas Murray (1987) and Murray et al. (1996) used the department as a whole as the unit of analysis. While the repercussions of this methodological difference are not totally clear, it seems unlikely that it is responsible for the significant longitudinal gains found by Murray, because the use of the department as the unit of analysis would tend to reduce sample size and decrease the probability of obtaining significant results.

An important limitation of the data in Figure 2 is that annual mean rating scores are based on a sample of teachers that varied somewhat from year to year due to faculty turnover. Thus, the possibility exists that year-to-year gains in mean student ratings were due not to longitudinal improvement in a fixed group of teachers (improvement by development), but rather to a tendency for newly appointed faculty members to

be better teachers, on average, than the individuals they replaced (improvement by selection). To check on this possibility a subsample of ten faculty members was identified who had held positions in the department for 26 consecutive years and had been evaluated by students in undergraduate courses in at least 22 of 26 years. Data for missed years (of which there were never more than two in succession) were estimated by interpolation. Figure 3 shows annual mean student rating scores for the fixed group of ten veteran faculty members and for the department as a whole.

Statistical analysis indicated that, despite small sample size, the fixed group of teachers showed significant longitudinal improvement over the 26-year observational period, but the amount of improvement shown by this group was significantly less than that for the department as a whole. The correlation coefficient between year and mean student rating was 0.74 for the fixed group of teachers, as compared to 0.91 for the department as a whole. These results indicate that the longitudinal gains in teacher ratings depicted in Figure 2 are due in part to true development over time in individual teachers and in part to the tendency of new faculty members to be more effective teachers than the individuals they replace.

Another question that arises from the data in Figure 2 is whether similar longitudinal improvement in student instructional ratings has occurred in other University of Western Ontario departments where teaching is evaluated in the same way as in the Department of Psychology. Figure 4 shows annual student evaluation data for

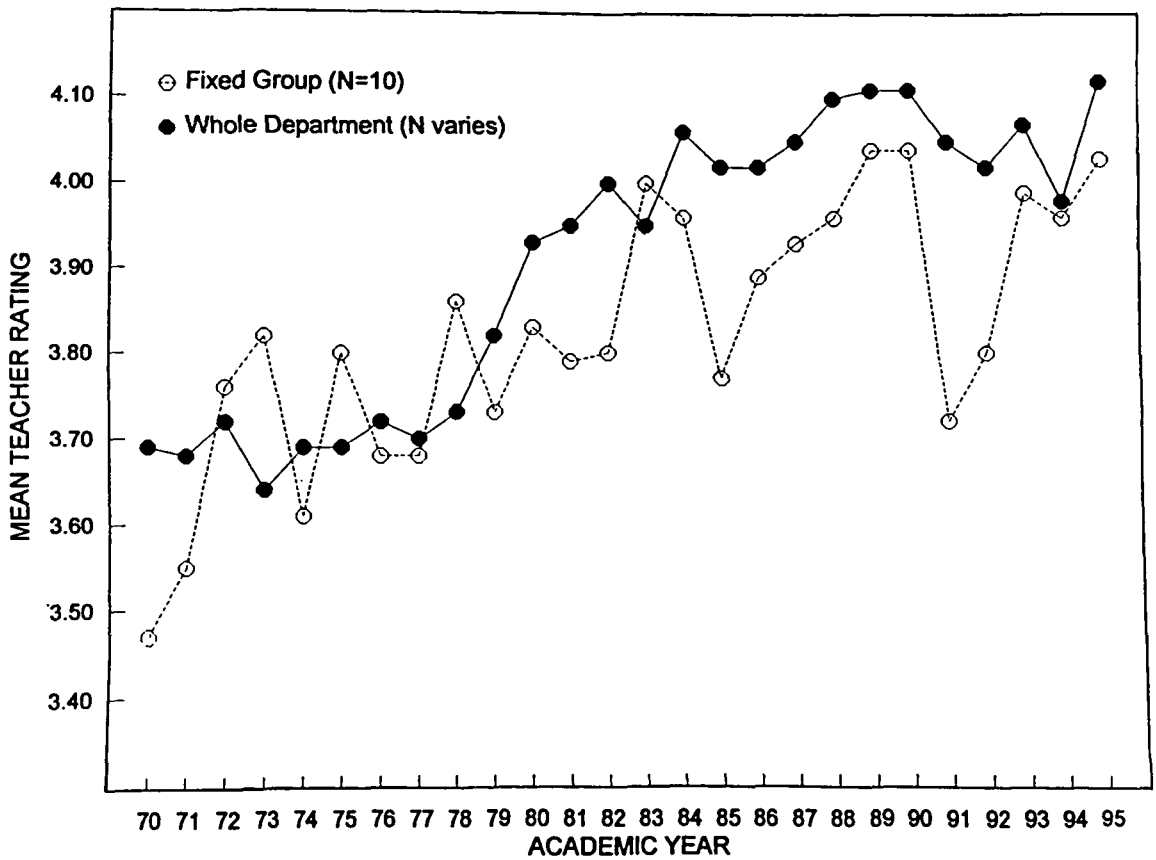

Figure 3 *Mean teacher rating scores for fixed group of faculty and for Department of Psychology as a whole for academic years 1969–70 to 1994–95.*

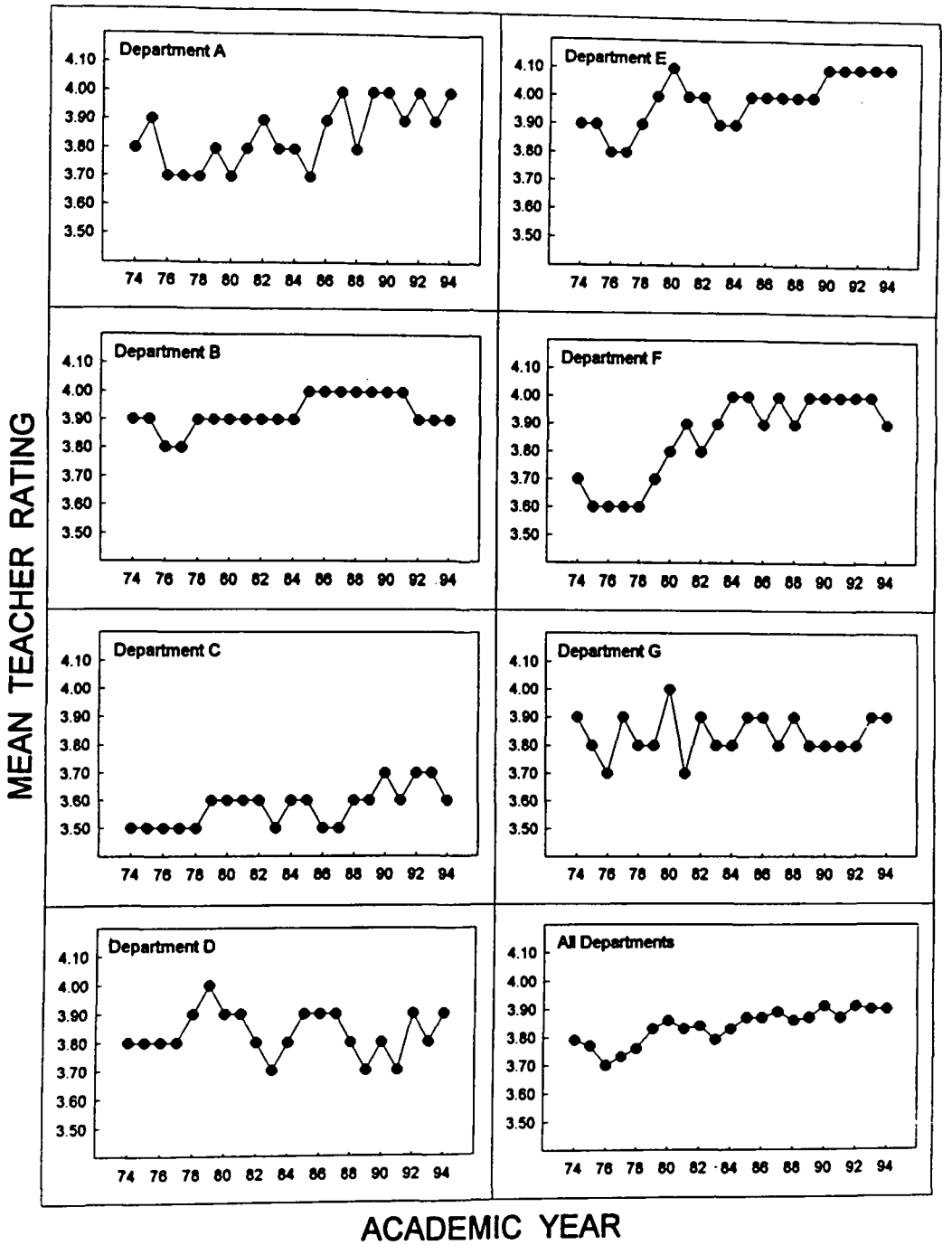

Figure 4 *Mean teacher rating scores for Faculty of Social Science departments for academic years 1973–74 to 1993–94.*

seven different departments in the Faculty of Social Science: Anthropology, Economics, Geography, History, Political Science, Psychology and Sociology; and for the Faculty as a whole. Departments are designated by letters in Figure 4 for protection of confidentiality.

Each of these departments uses the same ten-item teaching evaluation form as the Department of Psychology, and each administers this form annually in all courses for promotion and tenure purposes. Data are shown only for the years 1974 to 1994 inclusive, as these were the only years for which complete data were available for all departments. It may be noted that some departments showed clear longitudinal improvement in student ratings, whereas others showed no obvious change or perhaps even a slight decrease across years. Statistical analysis indicated that five of seven departments showed significant longitudinal gains in mean student ratings. The largest amount of improvement was found in Department F, but significant improvement was also evident in Departments A, B, C, and E. The correlation between year and student evaluation of teaching varied from +0.85 in Department F to +0.11 in Department G and -0.10 in Department D. These results suggest that it is possible to get conflicting longitudinal results even among similar academic units in the same institution that use the same teaching evaluation form. Thus the discrepant results of previous studies, and in particular the differing results of large-scale studies reported by Marsh and Hocevar (1991) and Murray et al. (1996), are perhaps not so surprising. But what is not clear are the reasons for these discrepancies.

What factors are responsible for finding long-term improvement in rated teaching effectiveness in some academic units but not in others? Given the pivotal role of expert consultation in field experiments reviewed earlier in this paper, it is plausible that participation in instructional development programmes, such as workshops, mini-courses, and peer consultation, might be one of the factors that contribute to long-term improvement in some academic units but not in others. Another possibility is that longitudinal improvement depends on the amount of weight placed on evaluation of teaching in decisions on faculty salary, promotion and tenure in a given faculty or department. These are interesting questions that invite further research.

In summary, research evidence indicates that introduction of student evaluation of teaching in an academic unit sometimes is, and sometimes is not, followed by improvement over time in mean student ratings for the unit as a whole. As noted previously, about 50% of studies in this area have reported significant longitudinal improvement and about 50% have reported no improvement. This indicates that long-term improvement in teaching following the introduction of student evaluation occurs under some conditions but not others. Unfortunately, we are not yet able to define the conditions that either enhance or inhibit performance.

Does evaluation of teaching lead to improvement of teaching?

Evidence reviewed above from three independent areas of research, namely faculty surveys, field experiments and longitudinal comparisons, suggests that student evaluation has indeed contributed to improvement of college and university teaching. Although each of these areas of research has its own methodological limitations, and each has yielded findings that show some degree of inconsistency, the convergence of evidence from three independent sources provides a strong case for concluding that, at least under certain conditions, student evaluation of teaching does lead to improvement of teaching. This conclusion, in combination with previous research demonstrating the reliability and validity of student evaluation of teaching, provides strong justification for the use (or continued use) of student instructional ratings as one of several types of information used in evaluation of faculty performance.

One interpretation of the research reviewed above is that student evaluation of teaching contributes to improvement of certain aspects of postsecondary teaching only, namely those aspects of teaching that are measured by the typical student evaluation form (eg, clarity of explanation, encouragement of student participation and promptness of feedback). It is possible, however, that other aspects of teaching have not benefited from student evaluation, and have actually gone in the opposite direction (ie in fact become worse) as a result of student evaluation of teaching. Consistent with this line of thought, it has been

argued that student evaluation causes grade inflation and lowering of academic standards together with entrenchment of traditional, outmoded styles of teaching. These possibilities are considered below.

Does student evaluation of teaching cause grade inflation?

One of the most frequent criticisms of student instructional ratings is that their mandatory use in faculty promotion and tenure decisions causes faculty members to raise grades and lower academic standards in an attempt to 'buy' positive ratings from students. Although this criticism is frequently raised, it is difficult to confirm or deny through research evidence. There are at least two types of research that are relevant to this issue, but in each case the evidence can be interpreted in more than one way. First, studies by Feldman (1976) reported correlations ranging from -0.04 to +0.63 (average = 0.28) between mean grades assigned by teachers (or expected by students) and mean student ratings received by the same teachers.

One possible interpretation of the correlation between teacher grades and student ratings is that teachers and students are involved in a 'mutual reinforcement process' whereby high grades from teachers are rewarded by high ratings from students. It is certainly plausible in this context that student evaluation of teaching could lead to grade inflation. On the other hand, the average correlation of 0.28 found between grades and ratings may reflect a tendency for highly rated teachers to foster high levels of learning in their students, which in turn results in justifiably higher student grades. In other words, the positive correlation between grades and ratings may be a valid reflection of differential teacher effectiveness rather than an impetus for grade inflation. If anything, research evidence supports the second rather than the first interpretation of the grades-ratings correlation (eg, Howard & Maxwell, 1980). It should also be noted that other research on student ratings (eg, Cashin, 1995) indicates a similar positive correlation of 0.20 to 0.30 between mean student rating of teaching and mean student rating of course difficulty. According to the first interpretation, this correlation might be taken to mean that teachers can 'buy' positive evaluations from their students by increasing, rather than decreasing, their course requirements and academic standards!

Surveys of faculty opinion provide a second line of research evidence relevant to the issue of whether student evaluation leads to lowering of academic standards. Table 2 summarizes the results of seven faculty opinion surveys that asked one or both of the following questions: 'Has student evaluation of teaching led to grade inflation?' and 'Has student evaluation of teaching caused lowering of academic standards?'

It is clear that results varied considerably from study to study. For example, support for the view that student evaluation contributes to grade inflation ranged from 14% in the McCready (1981) study to 87% in the Ryan, Anderson, and Birchler (1980) study. Across all studies reviewed in Table 2, and with differential weighting according to sample size, 27% of respondents said that student evaluation had caused grade inflation and 32% said that student evaluation had led to lowering of academic standards. It is cause for concern that approximately one-third of faculty members believe that student evaluation of teaching has had undesirable side-effects on academic standards. On the other hand, approximately two-thirds of faculty members believe that student evaluation does not lead to grade inflation or lowering of standards. Of course, there is the possibility that the results shown in Table 2 have been biased one way or the other by low return rates and/or by the invalidity of self-report. With respect to self-report validity, it is interesting to note that at least one faculty survey (Owens, 1977) found that the percentage of respondents reporting that their own grading standards had been affected by student evaluation was considerably less than the percentage who stated that their colleagues' grading had been influenced by student evaluation!

In summary, research evidence provides no clear support for the claim that student evaluation of teaching has led to grade inflation and lowering of academic standards. Although grade inflation does seem to have occurred in many institutions, and its occurrence may have coincided with the introduction of student evaluation of teaching, it is not clear that this was caused by student evaluation. As a case in point, if the use of student evaluation of teaching in faculty personnel decisions is one of the main causes of grade inflation, it is difficult to understand why grade inflation is apparent both in institutions where

Table 2 *Surveys of faculty opinion on undesirable side-effects of student instructional ratings*

Survey	Number of respondents	Percent Agreement	
		Have student ratings caused grade inflation?	Have student ratings led to lowering of standards?
Outcalt (1980) California, USA	4468	22	
McCready (1981) Wilfrid Laurier, Canada	25	14	
Gross & Small (1979) George Mason, USA	163	64	33
Murray et al. (1982) Western Ontario, Canada	666	27	
Ryan, Anderson, & Birchler (1980) Wisconsin-Lacrosse, USA	193	87	57
Dent & Nicholas (1980) Southern California,USA	120	59	
McMartin & Rich (1979) California State, USA	468	25	21

student evaluation is mandatory in promotion and tenure decisions and in institutions where it is not mandatory. Similarly, it is difficult to understand why grade inflation appears to be more pronounced in high schools than in colleges and universities, despite the fact that student evaluation of teaching is rarely if ever used for personnel evaluation purposes in high schools.

Does student evaluation of teaching lead to entrenchment of traditional methods?

A second criticism of student evaluation of teaching, the main proponent of which is Wilson (1987), is that student evaluation serves to maintain traditional teacher-centred methods of teaching (for example, the lecture method) and thus impedes instructional innovation and improvement. Wilson contends that the typical student evaluation form in current use implies a hierarchical, dictatorial, teacher-centred style of teaching. This is seen, for example, in items such as 'explains clearly', 'identifies important ideas', 'discusses alternative points of view' and 'motivates students to do their best work', all of which convey the idea that the education of students is the sole responsibility of the teacher rather than being shared equally between teacher and students. While not denying that student evaluation has led to the improvement of certain aspects of teaching, Wilson argues that the success and widespread acceptance of student evaluation has led to the entrenchment of 'restrictive and unjust' teacher-centred methods and the failure to develop innovative student-centred and shared-inquiry methods.

This is an interesting and provocative argument and, like the grade inflation issue, one that is not easy to resolve through empirical research. There are, however, several plausible counterarguments to Wilson's position. For one thing, contrary to what Wilson assumes, it is not totally clear that teacher-centred methods are inherently 'bad' and student-centred methods inevitably 'good'. It could be argued that each approach is appropriate in certain contexts and that some sort of balance between teacher- and student-centred methods is the ideal. Second, it is not clear that traditional

teacher-centred methods, such as lecturing, have increased in acceptability or credibility since the advent of student evaluation of teaching, as seems to be implied by Wilson's position. If anything, it would seem that teacher-centred approaches are used *less* today than was the case prior to the advent of student evaluation, whereas student-centred methods such as cooperative learning, discussion and problem-based learning are used *more* today. Similarly, and again in contradiction to Wilson's position, the use of innovative, student-centred methods of teaching appears to be more frequent in teachers who receive high ratings from students than in teachers who receive low ratings. Finally, it can be argued that student evaluation forms in current use are a reflection of what faculty already believe about teaching and not necessarily a prescription for what teachers should believe. The point is that we are free to construct new and different teaching evaluation forms if we think this is appropriate, and indeed some academic units have developed alternative evaluation forms for teacher-centred courses.

In summary, although data are limited, there are no strong reasons for believing that student evaluation of teaching perpetuates traditional instructional methods and impedes innovation. It appears that teacher-centred methods and resistance to innovation were well-established in higher education long before the advent of student evaluation, and continue to be well-established today.

Conclusions

1. Converging evidence from three independent sources, namely faculty surveys, field experiments and longitudinal comparisons, suggest that student evaluation of teaching has contributed to improvement of certain aspects of college and university teaching.
2. The contribution of student evaluation to the improvement of teaching is greatly enhanced by expert consultation with instructional development specialists. This finding provides support for the positive impact of instructional development offices and programmes in improving teaching. More research is needed to decide the most effective ways of combining student evaluation with expert consultation.
3. There is no clear evidence that student evaluation of teaching has led to negative side-effects commonly attributed to it, such as grade inflation and entrenchment of traditional methods of teaching.
4. Evidence that student evaluation leads to significant improvement of teaching, in combination with research demonstrating the reliability and validity of student evaluation forms, provides strong justification for the use of student evaluation of college and university teaching, both as diagnostic feedback to faculty members and as one of several sources of information considered in decisions on faculty hiring, retention, salary and promotion. However, since students are capable of assessing only some aspects of teaching, student evaluation should never be the only source of data on teaching in faculty personnel decisions.

References

Baker, A. M. (1986). Validity of Palestinian university students' responses in evaluating their instructors. *Assessment and Evaluation in Higher Education, 11,* 70–75.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited.* Idea Paper No. 32. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J. A. (1973). Effectiveness of student feedback in modifying college instruction. *Journal of Educational Psychology, 65,* 395–401.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education , 13,* 321–341.

Cohen, P. A. (1981). Student ratings of teaching and student achievement: A meta-analysis of multi-section validity studies. *Review of Educational Research, 51,* 281–309.

Dent, P. L., & Nicholas, T. (1980). A study of faculty and student opinions on teaching effectiveness ratings. *Peabody Journal of Education, 26,* 135–137.

Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education, 4,* 69–111.

Gray, D. M., & Brandenburg, D. C. (1985). Following student ratings over time with a catalog-based system. *Research in Higher Education, 22,* 155–168.

Gross, R. B., & Small, A. C. (1979). A survey of faculty opinions about student evaluations of instructors. *Teaching of Psychology, 6,* 216–219.

Howard, G. S., & Maxwell, S. E. (1980). The correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology, 72,* 810-820.

Jacobs, L. C. (1987). *University faculty and students' opinions of student ratings.* Bloomington, IN: Bureau of Evaluative Studies and Testing. (ERIC Document Reproduction Service No. ED 291 291)

Marsh, H. W, & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research.* Volume 8. New York: Agathon.

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,* 303-341.

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,* 217-251.

McCready, D. J. (1981). *Student evaluations of teaching* (Research Paper Series No. 8017). Waterloo, Canada: Wilfrid Laurier University, School of Business and Economics.

McKeachie, W. J., Lin,Y. G., Daugherty, M., Moffet, M. M., Nork, J., Walz, M., & Baldwin, R. (1980). Using student ratings and consultation to improve instruction. *British Journal of Educational Psychology, 50,* 168-174.

McMartin, J. A., & Rich, H. E. (1979). Faculty attitudes toward student evaluation of teaching. *Research in Higher Education, 11,* 137-152.

Menges, R. J. (1980). Student evaluations of instruction and faculty morale. *Instructional Evaluation, 5* (1), 16-18.

Menges, R. J., & Brinko, K. T. (1986, April). *Effects of student evaluation feedback: A meta-analysis of higher education research.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Miller, A. H. (1988). Student assessment of teaching in higher education. *Higher Education,17,* 3-15.

Murray, H. G. (1980). *Evaluating university teaching : A review of research.* Toronto, Canada: Ontario Confederation of University Faculty Associations.

Murray, H.G. (1987a). Acquiring student feedback that improves instruction. In M. G. Weimer (Ed.), *Teaching large classes well.* New Directions for Teaching and Learning. No. 32. San Francisco: Jossey-Bass.

Murray, H. G. (1987b, April). *Impact of student instructional ratings on quality of teaching in higher education.* Paper presented at the annual meeting of the American Educational Research Association, Washington. (ERIC Document Reproduction Service No. ED 284 495)

Murray, H. G., Jelley, R. B., & Renaud, R. D. (1996, April). *Longitudinal trends in student instructional ratings.* Paper presented at annual meeting of the American Educational Research Association, New York.

Murray, H. G., Newby, W. G., Bowden, B., Crealock, C., Gaily, T. D., Oswin, J., & Smith, P. (1982). *Evaluation of teaching at the University of Western Ontario.* London, Canada: University of Western Ontario, Provost's Advisory Committee on Teaching and Learning.

Murray, H. G., & Smith, T. A. (1989, March). *Effects of midterm behavioral feedback on end-of-term ratings of instructor effectiveness.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Ory, J. C., & Braskamp, L. A. (1981). Faculty perceptions of the quality and usefulness of three types of evaluative information. *Research in Higher Education, 15,* 271-282.

Outcalt, D. L. (1980). *Report of the task force on teaching evaluation.* Santa Barbara, CA: University of California.

Overall, J. U., & Marsh, H. W. (1979). Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology, 71,* 856-865.

Owens, R. E. (1977). *Evaluating the importance of teaching.* Manhattan, KS: Kansas State University, Office of Educational Research. (ERIC Document Reproduction Service No. ED 160 016)

Pigott, A., & Rosehart, R. G. (1983). Development and use of student course evaluations at Lakehead University. In J. G. Donald (Ed.), *Proceedings of the Montebello Conference on the Evaluation and Improvement of University Teaching: The Canadian Experience.* Montreal, Canada: McGill University, Centre for University Teaching and Learning.

Prosser, M., & Trigwell, K. (1990). Student evaluation of teachers and courses: Student study strategies as a criterion of validity. *Higher Education, 20,* 135-142.

Ryan, J. J., Anderson, J. A., & Birchler, A. B. (1980). Student evaluation: The faculty responds. *Research in Higher Education, 12,* 317-333.

Stevens, J. J., & Aleamoni, L. M. (1985). The use of evaluative feedback for instructional improvement: A longitudinal perspective. *Instructional Science, 13,* 285-304.

Vogt, K. E., & Lasher, H. (1973). *Does student evaluation stimulate improved teaching?* Bowling Green, OH: Bowling Green State University, College of Business. (ERIC Document Reproduction Service No. ED 078 748)

Watkins, D. (1994). Student evaluations of university teaching: A cross-cultural perspective. *Research In Higher Education, 35*, 251–266.

Wilson, R. C. (1986). Improving faculty teaching: Effective use of student evaluations and consultants. *Journal of Higher Education, 57*, 196–211.

Wilson. T. C. (1987, April). *Pedagogical justice and student evaluation of teaching forms: A critical perspective.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

The author

Harry Murray is Professor of Psychology at the University of Western Ontario in London, Canada. He earned BA and MA degrees from Western Ontario, plus a doctoral degree in experimental psychology from the University of Illinois in 1968. His major areas of research interest are teacher effectiveness and teaching evaluation in higher education. Academic honours include a 3M Canada fellowship for outstanding teaching and the W J McKeachie Career Achievement Award from the American Educational Research Association.

Address: Department of Psychology, University of Western Ontario, London, Ontario, Canada N6A 5C2.
Tel: (519) 661 2067; Fax: (519) 661 3961
E-mail: murray@vaxr.sscl.uwo.ca