Student Evaluation of Teaching: Has It Made a Difference?

Paper presented at the Annual Meeting of the
Society for Teaching and Learning in Higher Education

Charlottetown, Prince Edward Island, June 2005

Harry G. Murray
Department of Psychology
University of Western Ontario

Preliminary Comments

I want to start out by thanking the Society for Teaching and Learning in Higher Education for the Christopher Knapper Award, and to thank Chris personally for his support over the years.

When I first started doing research in the area of college and university teaching about 35 years ago, it was like being a voice in the wilderness. But I discovered two other voices in the same wilderness doing the same type of research, namely Chris Knapper and Bill McKeachie. I have benefited greatly from my contacts with both of them over the years, and I have also been lucky enough to have won awards named after both of them! I guess when you start winning career achievement awards such as the Chris Knapper Award, you know you are getting old, but on the other hand, when you have awards named after you, as Chris does, you know you are getting really old!

My topic for today is student evaluation of teaching in colleges and universities. This is a practice that has been going on for 35 or 40 years in North American institutions, and in my view, is one of the most controversial and most interesting developments in higher education in recent times.

Much of the research done on student evaluation of teaching has focused on the issue of reliability and validity. Do student evaluations of teaching provide replicable and accurate information about quality of teaching? I will comment briefly on reliability and validity, but will focus mainly on a less-researched issue, namely the impact of student evaluation of teaching. Has it made a difference, either positive or negative? Are we better off or worse off as a result of student evaluation of teaching?

History and Rationale of Student Evaluation of Teaching

Before getting to the impact of student evaluation of teaching, I want to spend a few minutes on the history of student evaluation, and the rationale of student rating forms.

In most colleges and universities in North America, student evaluation of teaching began in the late 1960's or early 1970's. The earliest use of student evaluation of teaching that I know of was at University of Washington in the 1920's, initiated by psychologist E.T. Guthrie. At my university, the University of Western Ontario, student evaluation began in the late 1960's, and was supported by a coalition of three groups:

- students who wanted a say in teaching;
- administrators who were concerned with accountability and good public relations (i.e., we are doing something about teaching);
- young faculty who wanted their salary, promotion and tenure evaluations to depend on something other than number of publications alone.

There was a lot of opposition and lots of controversy regarding student evaluation of teaching, as there still is. One incident I recall from this time period is that of a Dean—an eminent economist—who was initially opposed to student evaluation of teaching, but changed his mind when he got a 1.4 rating and thought a >1 was the top rating. We didn't have the courage to tell him that 1.4 was actually an

extremely low rating! And I believe he went back to being opposed to student evaluation when he found out the real meaning of 1.4. Despite the controversy, student evaluation of teaching got accepted and spread like wildfire across North America and other countries. Today nearly 100% of North American institutions make use of student evaluations, and according to surveys I have seen, something like 70-75% of faculty members support the use of student evaluations. And we all know that getting 75% of faculty members to agree on anything is a big accomplishment!

On most campuses, student evaluation of teaching is done by means of a brief, standardized rating form on which student's rate characteristics of teachers and courses, such as clarity of explanation, enthusiasm, availability, and fairness of exams, usually on a 5-point rating scale. Although these forms are widely used, it might be a good idea to remind ourselves of the underlying rationale of student rating forms, and the inherent limitations of such forms.

Student rating forms are really intended as a substitute or proxy for direct measurement of student learning, which most people would consider to be the best way to measure teaching, but something that is fraught with technical difficulties. As shown in **Slide 1**, student rating forms try to do the next best thing by assessing teacher or course characteristics that are:

a. believed to contribute to student learning, based on evidence or logical argument;
b. observable by students;
c. widely applicable, and thus can be used in many different courses; and,
d. under the control of the instructor, and thus are justifiable for use in faculty personnel decisions on salary, promotion and tenure.

These four item selection criteria are necessary for a student rating form intended for use in salary, promotion, and tenure decisions, but application of these criteria severely limits the range of teacher and course characteristics that can be included, and thus impose inherent limitations on student evaluation of teaching. For example, because student evaluation forms can assess only those characteristics that are observable by students, they cannot assess non-classroom factors such as course design, and substantive factors such as instructor knowledge, academic standards, and quality of assignments. Similarly, characteristics such as quality of handouts and coordination of lab work, although potentially useful as feedback to the instructor, are applicable only in some courses and thus not appropriate for a standardized rating form. Thus we must admit at the outset that student evaluation of teaching is incomplete and lacking in scope, and must always be supplemented by other sources of data on teaching.

Are Student Ratings Valid?

Now let me briefly review research on the reliability and validity of student evaluation of teaching. How well do student ratings do in providing a reliable and valid assessment of quality of teaching? This question has received a lot of attention in the research literature, with over 2,000 published studies!

Research indicates that student ratings are adequate in terms of reliability, in that ratings of a given instructor are reasonably stable or consistent across courses, years, rating forms, and groups of raters. Other research indicates that student evaluations are valid or accurate in that they are relatively free of bias, and agree with evaluations made by others, such as colleagues and alumni. But in my view, the most important evidence on validity comes from the two types of studies described below. First are classroom observation studies in which trained observers visit classes to record the frequency with which instructors exhibit specific, "low-inference" teaching behaviours such as "signals the transition from one topic to the next" and "addresses individual students by name." Then an attempt is made to predict student ratings of teaching from outside observer reports of specific teaching behaviours. As illustrated in Slide 2, this type of research has shown that student ratings are closely related to and highly predictable from specific classroom behaviours of the instructor. Slide 2 shows sample results for six teaching behaviours selected from a total of 50 studied by Murray (1985).

The multiple correlations between the first ten classroom behaviours selected in stepwise multiple regression and end-of-term student instructional ratings was approximately .90 in this study, with the

highest individual correlations observed for teaching behaviours related to clarity, expressiveness, and interaction. In other words, student ratings can be viewed as "valid" in that they reflect what the instructor actually does in the classroom, rather than by irrelevant factors such as "popularity" or "personal warmth."

The second type of evidence supporting the validity of student ratings comes from what are called "multi-section validity studies." These are field studies of courses with many different class sections taught by different instructors, but with a common syllabus and common final exam that is objectively scored, either by computer or by an exam marking committee. In this situation it is assumed that differences in section mean scores on the common final exam reflect differences in amount learned by students in the various class sections, rather than simply differences in instructor grading practices. So, the key validity question is whether instructors who get high ratings from students do in fact teach their students more effectively so that they do better on the common final. Slide 3 shows the average or typical result found in studies of this type. This is a scatter plot showing the correlation across class sections between mean student rating of teaching and mean score on the common exam. Each of the dots represents one teacher or class section. The typical result in these studies is a correlation of approximately .50 (Cohen, 1981). You can see in the scatter plot that this degree of correlation reflects a general trend whereby teachers with higher ratings tend to have students who do relatively well on the common final. Thus, students taught by highly rated teachers tend to learn the subject matter better than those taught by lower rated teachers, or putting it in another way, student ratings validly reflect differences in actual teaching effectiveness, rather than extraneous variables. On the other hand, it is obvious that student ratings are not a perfect indicator of differences in teacher effectiveness, as indicated by exceptions from the general trend of the scatter plot in which exam scores are low for highly rated teachers, or exam scores are relatively high for low rated teachers. Again, we see that if we rely solely on student ratings as a measure of teaching effectiveness, we will inevitably make some mistakes.

There are two other points worth making about the results of the multi-section validity studies. First, although the average correlation found in these studies is approximately .50, the correlation found in individual studies varied widely, all the way from -.70 to +.90, although all studies but one yielded a positive correlation. The reasons for this variability are not clear, but it illustrates the contextuality of research findings on student evaluation of teaching. What is found seems to depend a lot on context. For example, the correlation seems to be higher in some academic disciplines, possibly higher in mathematics than in social science. The highest correlation was .90 in a first year Calculus course. Second, why is the average correlation only .50? Why isn't it higher? I can think of two possible reasons. First, the types of common final exams used in multi-section validity studies, namely multiple-choice or fact recall exams, may not be the type that best reflects the instructor's contribution to student learning. We might get a higher correlation if we measured student learning by exams that focus on student problem solving, critical thinking, or organization of concepts. Second, although student ratings appear to be a valid measure of classroom teaching, it seems obvious that classroom teaching is only one of several teacher and course characteristics that contribute to student learning. Students are not in a good position to judge factors such as instructor knowledge, quality of readings, course management, and academic standards used in grading, but these factors may have a strong impact on student learning. Factors such as these could potentially be assessed by colleagues, and if we used student evaluation of classroom teaching in combination with colleague evaluation of substantive and non-classroom aspects of teaching, we would probably find a correlation of larger than .50 with measures of student learning.

Impact of Student Evaluation of Teaching

Turning now to my main topic, the impact of student evaluation of teaching, I will discuss whether student evaluation has made a difference in three areas: faculty personnel decisions, improvement of quality of teaching, and academic standards.

Impact on Personnel Decisions

Does student evaluation of teaching make a difference in faculty personnel decisions? Are student ratings taken into account to any significant extent in decisions on faculty tenure, promotion, retention, and salary increments? This was one of the stated purposes of student evaluation of teaching when evaluation was first introduced, so it is of interest to determine if this purpose has been achieved.

Slide 4 summarizes the three types of research designs that have been used in studying this question (Murray, 1984). First are surveys of faculty opinion, in which faculty members are asked to estimate, in percentage terms, the amount of weight they believe is placed on student evaluation of teaching relative to research and service work in salary, promotion and tenure (SPT) decisions. The typical result is an estimate from faculty that student evaluation of teaching is weighted around 20-30% in SPT decisions, compared to 60% or more for research. Other findings from this research are as follows:

a. Teaching is seen to be weighted higher in annual salary adjustments than in promotion and tenure decisions
b. The weight placed on teaching varies across academic disciplines. For example, a survey at the University of Western Ontario indicated higher weightings in professional schools than in traditional Arts and Science faculties.
c. Most faculty want more weight placed on teaching than is presently the case.

The second type of research design used to assess the impact of student evaluation on personnel decisions is field studies of actual SPT decisions in a university department. Given that you know the outcome of such decisions for a sample of faculty members, and you have data on teaching and research performance for these same faculty members, you can use regression methods to estimate the contribution of teaching and research to SPT decisions. Not many studies of this type have been done, but the usual finding is that approximately 10% of the variance in SPT outcomes is accounted for by student evaluation of teaching. Another finding of these studies is that student ratings of teaching show a correlation of zero with measures of research such as number of publications, contrary to the view that teaching and research productivity are mutually complementary.

The third type of study could be called a simulation study. Here the researcher experimentally constructs personnel files for hypothetical faculty members, with variation in both teaching and research assessments (e.g. high student evaluation of teaching, low number of publications or vice versa). The files are then given to people who serve on Promotion and Tenure Committees and they are asked what decision they would make regarding salary, promotion or tenure. The general conclusion from studies of this type is the same as that reported above, namely teaching does have an impact on salary, promotion and tenure decisions, but it is a small impact compared to that of research. Another finding from these studies is that if statistical summaries of student evaluation of teaching are supported by prose comments from students, they have a greater impact on salary, promotion and tenure. So good numerical ratings supported by positive prose comments are more likely to lead to a positive decision than ratings alone, whereas poor ratings accompanied by negative comments are more likely to lead to a negative decision. It appears that the prose comments make teacher evaluation data more vivid or convincing or meaningful, and this works in both directions.

In summary, research indicates that student evaluation of teaching does make a difference in personnel decisions, but not a large difference relative to research.

Impact on Quality of Teaching

Now we come to what I would consider to be the most important question of all: Has student evaluation of teaching led to actual improvement in quality of teaching?

I think this question is most important because, as with all forms of evaluation, improvement of performance is, or is supposed to be, the ultimate purpose of teaching evaluation, and the most justifiable reason for doing evaluation in the first place.

My impression, for what it is worth, is that teaching in universities is of better quality today than it was 30 or 40 years ago, and that student evaluation of teaching has contributed to this improvement. My belief that teaching has in improved is supported by the results of a survey of senior faculty members at the University of Western Ontario conducted by one of my students, Linda Dash, as a senior honours thesis (Dash, 1992). In this survey, 68% of respondents said that university teachers of today are more skilled than when the respondent began his or her university career 30 or more years ago, whereas 25% said today's teachers are worse, and 7% were undecided.

Assuming that improvement has in fact taken place, is there any research evidence to support the view that student evaluation of teaching has contributed to this improvement? Again, as indicated in Slide 5, I will briefly review research evidence from three sources (Murray, 1997).

Faculty Opinion Surveys

Slide 6 shows the average result of eight surveys of faculty opinion in which the following two questions were included:

1. Do student evaluations of teaching provide useful feedback for improvement of teaching?
2. Have student evaluations of teaching led to improved teaching?

As can be seen, 73.4% of faculty agreed that student evaluation of teaching provides useful feedback, and 68.8% agreed that student evaluation of teaching had improved teaching. Thus, the majority of faculty members agreed that student evaluation of teaching had led to improved teaching. To be fair it must be acknowledged that one or two faculty opinion surveys have revealed very negative attitudes toward student evaluation of teaching (e.g., Ryan, Anderson, and Birchler, 1980). As with most research done in this area, the results of faculty opinion surveys are not totally consistent across studies.

Field Experiments

The second source of research evidence is from field experiment, done with actual college or university teachers or teaching assistants, in which, as shown in Slide 7, an experimental group of teachers receives feedback from student evaluation of teaching done at mid-term, whereas the control group is evaluated but does not receive mid-term feedback. The two groups are then compared on end-of-term student evaluations. If student evaluation contributes to improved teaching, we would expect the experimental group to benefit from mid-term feedback and show a larger gain in end-of-term ratings than the control group. Slide 8 summarizes the results of 22 field experiments reviewed by Cohen (1980). It may be noted that the mean gain due to feedback for the experimental group was .10 points on a 5-point scale (3.70 to 3.80) when feedback consisted of numerical student ratings alone, but the gain was much larger, .33 points (3.70 to 4.03) when student feedback was accompanied by consultation with a faculty development expert, who assisted in interpreting the student ratings and provided specific suggestions for improvement. The gain due to student feedback alone corresponds to a gain of 8 percentile points (e.g. 50th to 58th percentile), whereas that for feedback plus consultation corresponds to a gain of 24 percentile points (e.g., 50th to 74th percentile), which is a very large and significant gain in practical terms.

The field experiments provide further support for the conclusion that student evaluation of teaching can lead to improved teaching, but the amount of improvement in these studies is not large unless student feedback is accompanied by expert consultation. This latter finding gives support to the important contribution made by faculty development offices to quality of teaching. It appears from these data that student evaluation of teaching and faculty development programs play complementary and synergistic roles in teaching improvement. Student evaluation of teaching increases the need or demand for faculty development programs, whereas faculty development programs provide an avenue for translating student feedback into actual improvement of teaching.

Longitudinal Comparisons

The third type of research evidence on the question of whether student evaluation of teaching leads to improvement of teaching comes from studies in which student ratings in a given academic unit (e.g. department or faculty) are compared or tracked longitudinally over time, beginning at the point where student evaluation of teaching was first introduced in that unit. If evaluations contribute to improvement of teaching, then you would expect to see gradual longitudinal improvement in the *average* teacher rating score in an academic unit. I know of 14 studies that have reported longitudinal comparisons of this sort, and the results of these studies again are somewhat inconsistent: 8 found improvement across successive years, and six did not. It is important to note that most of these studies did not include the methodological conditions that are needed for optimal testing of longitudinal trends. For example, some did not start at Year 1 of the student evaluation of teaching program, some included only two to three

years of data, and some did not use the same student evaluation of teaching instrument throughout. However, one study conducted by myself and some of my students at the University of Western Ontario did have all the optimal conditions for longitudinal comparison. At UWO, mandatory student evaluation of teaching for all teachers and courses has been in effect since 1969-70, and the same student evaluation form was used for a 26-year period from 1970 to 1995. Slide 9 shows the average teacher rating for all full-time faculty members in the UWO Department of Psychology (N= 40 to 50 faculty members per year) in each of 26 consecutive academic years from 1969-70 to 1994-95. The reason the data end in 1995 is that we switched to a different student rating form with a 7-point rating scale at that time. It may be noted that significant longitudinal improvement did take place in this study. The department mean teacher rating increased from around 3.70 in 1969-70 to around 4.10 in 1994-95, an increase of more than one standard deviation (Murray, 1997).

One possible criticism of these data is that the improvement across years resulted at least in part from changes in the composition of the department. For example, it is possible that people who left the department in this time span tended to be poorer teachers, whereas those who replaced them were better (and younger) teachers. This criticism does not appear to be viable, however, because more or less the same trend across years as shown in Figure 3 was found when data were plotted only for a fixed group of faculty (N=10) who stayed in the department throughout the full 26-year period.

As stated previously, the results shown in Slide 9 were not always found in other studies, and one of the studies failing to get the predicted result, that of Marsh and Hocevar (1991), included most of the optimal methodological conditions needed for testing longitudinal trends. In addition, even within the University of Western Ontario study, some academic departments showed results similar to that for Psychology, whereas others did not. So once again we have a degree of inconsistency or contextuality that seems to be unavoidable in research on student evaluation of teaching.

In conclusion, when we put together the results of these three types of research, namely faculty opinion surveys, field experiments and longitudinal comparison, we have converging evidence that student evaluation of teaching has contributed to improvement of teaching, despite the fact that the improvement (1) is often not large in absolute terms, and (2) seems to occur more in some contexts than in others. So my conclusion, with some reservations, is that teaching has improved in colleges and universities, and student evaluation of teaching has contributed to that improvement.

Impact on Academic Standards

The third area where student evaluation of teaching is claimed to have had an impact is academic standards. In this case the impact is negative. A strong and persistent criticism of student evaluation of teaching is that it causes grade inflation and lowering of academic standards. It is claimed that since faculty members know that student evaluations are used in personnel decisions and are afraid that giving low grades to students will cause retribution from students in the form of low teacher ratings, they respond by raising grades, which leads to rampant grade inflation.

In support of this claim, research indicates that there is indeed a significant correlation between grades given and student ratings received. Teachers who give higher grades do in fact get higher student ratings. The correlation is around .30, on average, across studies. In addition, surveys of faculty members indicate that a sizeable minority of faculty (about 25-30%) believe that student evaluation of teaching do cause grade inflation. It is interesting to note that in some studies, faculty members believed that student evaluation of teaching caused <u>other</u> faculty members to inflate their grades, but had no influence on their own grading practices! But we must keep in mind that a correlation between grades and ratings does not necessarily imply a cause-effect relationship between grades and ratings. In fact, as illustrated in Slide 10, there are several possible underlying cause-effect patterns that could produce a correlation between grades and ratings. For example, the grade inflation hypothesis says that high or low grades do in fact cause students to give high or low ratings, such that student evaluation of teaching is a cause of grade inflation. On the other hand, the teacher effectiveness hypothesis says that more effective teachers tend to receive higher ratings from students, but at the same time tend to foster higher levels of learning in students, which is justifiably reflected in higher grades. Thus ratings are correlated with grades because both higher ratings and higher grades are both caused by the same thing, namely teacher effectiveness. So highly-rated teachers do give higher grades, but do so because their students are

actually learning more. Yet another possibility is that some other factor, such as course level, causes variation in both ratings and grades, and this is the reason that grades and ratings are correlated. For example, it may be students Year 4 course tend to both give higher ratings and earn higher grades than students in Year 1 courseS.

Much research has been done attempting to discover which of these alternative causal hypotheses is correct, but there appears to be no clear resolution. There are studies supporting both the grade inflation hypothesis and the teacher effectiveness hypothesis. One finding that supports the grade inflation hypothesis is that a relationship between grades and ratings is found in studies where student grades are experimentally manipulated with other variables controlled. An example of this was a recent study by Johnston (2003) and another by Worthington and Wong (1979) at Trent University. These studies were carried out despite ethical problems involved in experimental manipulation of student grades. On the other hand, contrary evidence comes from the fact that grade inflation is equally prevalent in institutions that do and do not use student evaluation of teaching (e.g., university vs. college vs. high school).

Even if it were shown that grading level is a cause or determinant of student evaluation of teaching, there is a solution to this problem, namely to adjust student evaluations of teaching upward or downward to take account of the biasing effect of strict or lenient grading, just as other potential bias factors can be controlled by using separate norms or statistical adjustments for different types of classes, for example, large vs. small classes and optional vs. required courses.

In summary, a lot of people believe that student evaluation of teaching has had an impact on academic standards, namely a negative impact involving inflation of grades and weakening of academic requirements, but research evidence is not clear in support of this belief.

Conclusions

Regarding the impact of student evaluation of teaching, I believe that it has had an impact, it has made a difference, and with some reservations, I believe the impact has been positive or beneficial. In particular, I believe that university and college teaching has improved over the past 30-40 years, and this improvement is partly due to student evaluation of teaching. I believe it is also partly due to the faculty development movement, which as stated earlier, I see as working synergistically with student evaluation of teaching. Second, I don=t believe there is clear evidence that student evaluation of teaching has had a negative impact on academic standards by causing grade inflation.

As I stated earlier, I have some reservations in making this conclusion, and my two main reservations are as follows. First, the possibility remains that student evaluations of teaching does cause grade inflation or are biased or unfair in other ways. We need to be continuously vigilant about such possibilities, and if we find biases or negative side-effects, take steps to eliminate them or correct them, for example, by statistical adjustment or grouping teacher ratings according to course type. My second reservation is even though student evaluation of teaching has been found to be reasonably good in terms of reliability and validity of measurement, and has also been found to contribute to improved teaching, we must not lose sight of the fact that student evaluation of teaching has inherent limitations, and thus we never should rely on student ratings alone in evaluation of teaching. Students can only evaluate what they can observe, and what they observe is mainly what occurs inside the classroom. But as stated previously, there are other very important components of teaching, such as course quality, instructor knowledge, quality of assignments, and curriculum development that cannot be measured by student ratings, and need to be assessed in some other way.

What I am saying is that we should continue to use student evaluations of teaching, but we should be looking for other, alternative methods of evaluating teaching that provide a necessary supplement to those instruments.

Possible Alternatives to Student Evaluations of Teaching

Slide 11 lists four alternative ways of evaluating teaching that could be used as supplements to student evaluation of teaching. One possibility is direct measurement of student learning. In other words,

we measure how much students have learned and evaluate teachers accordingly. The best teacher is the one whose students learn the most, and the worst teacher is the one whose students learned the least. This idea has a lot of intuitive appeal, but there are many problems and technical difficulties that make this plan next to impossible to implement. For example, we would need to construct "pre" and "post" tests for all courses, we would need to get instructors to specify their course objectives as a prerequisite to test construction, and we would have difficulty in comparing pre-post gain scores across different types of courses. This approach is probably only viable in situations where there is a widely accepted standardized test that assesses basic student knowledge in an academic discipline. Apparently there is such a test in Physics, the "Force Concept Inventory", but this is rare, and even with such a test, how do you connect test results directly to the efforts of individual teachers?

A second option is performance indicators. This is what *MacLean's* magazine uses, for better or for worse, in evaluating and ranking Canadian universities. Performance indicators are objective, readily available measures that supposedly reflect institutional performance or quality, For example *MacLean's* uses indicator variables such as number of books in library, percentage of budget allocated to scholarships, and percentage of Year 1 classes taught by tenured faculty. Performance indicators could potentially be used to evaluate individual teachers and courses rather than whole institutions, provided we are able to identify valid indicator variables that are appropriate for individual teachers. The main problem with performance indicators as used by *MacLean's* magazine is that no attempt has been made to demonstrate that the performance indicators are valid measures of educational quality. For example, as far as I know, there is no evidence that students learn more or are more satisfied with their education or become better life-long learners at institutions that have more books in the library or more Year 1 courses taught by tenured faculty. Performance measures may be reliable, but they have not been shown to be valid. Despite this, I am amazed that universities and university presidents seem to view the *MacLean=s* rankings with great respect and treat them as totally valid measures.

If performance indicators can be found that are appropriate for evaluating individual teachers or courses, we would then need to validate the indicators. One of my former graduate students, Robert Renaud, did this with respect to one potential performance indicator, namely the frequency of higher-order, thought questions asked on exams or assignments in a university course (Renaud, 2002). He found a significant correlation between frequency of higher-order questions on assignments and student pre-post gains in critical thinking ability.

A third possible alternative or supplement to student ratings is colleague evaluation of teaching. I believe this to be the potentially most useful of the four alternative methods listed in Table 6. As stated previously, students are not qualified to assess the more substantive or content-oriented aspects of teaching, such as instructor knowledge, academic standards, and exam quality, nor are they in a position to evaluate components of teaching that occur outside the classroom, such as course design, and curriculum development. Colleagues are perhaps in a better position to assess these factors and could probably do so through inspection of course materials and documents (e.g., course outline, exams, handouts) in a manner similar to what colleagues do in evaluating a faculty member's research contributions in relation to a promotion or tenure decision. Colleagues might also want to attend some classes to round out their evaluations, although students are already doing an adequate job of evaluating classroom teaching. If colleague evaluation of substantive and non-classroom components of teaching were combined with student evaluation of what happens inside the classroom, we would be a lot closer to a fully satisfactory teaching evaluation system, and in my opinion, evaluation of teaching would have more credibility and greater weight in personnel decisions. However, colleagues, for various reasons are reluctant to participate in evaluation of teaching, so the ideal teaching evaluation system is not in place at many institutions.

A fourth and final alternative method of evaluation is the *National Survey of Student Engagement.* The NSSE, which has received a lot of attention lately, is a questionnaire given to students in which they rate the extent to which they are actively involved or engaged in their education, for example by way of participation in class discussion, student-faculty interaction, independent projects, and community service. This is somewhat like traditional student rating forms for evaluation of teaching, except that students are evaluating themselves, and are evaluating activities that occur both inside and outside the classroom. In its present form, the NSSE is intended to evaluate institutions rather than individual

teachers, but is could possibly be modified to be appropriate for individuals. Based on what I have seen so far, it does have some potential as a measure of quality of education or quality of teaching.

Final Comments

The bottom line, in my view, is that we should keep student evaluation of teaching, but supplement it with one or more of the above four alternatives, preferably colleague evaluation, or with something better if it comes along. Research indicates that student evaluation of teaching is more than adequate in terms of reliability and validity, and has led to improvement of teaching. Student evaluation of teaching has value and is worth keeping, but it is a mistake to assume that student evaluation provides a complete assessment of all important aspects of college or university teaching. Student evaluation of classroom teaching in combination with colleague evaluation of substantive and non-classroom aspects of teaching comes much closer to telling the whole story.

References

Cohen, A.P. (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 instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research* 51:281-309.

Dash, L. (1992). *Survey of Faculty Opinion on Long-term Trends in Quality of Higher Education*. Unpublished honours thesis. University of Western Ontario, London, Ontario, Canada.

Johnson, V. E. (2003). *Grade Inflation: A Crisis in College Education*. New York: Springer-Verlag.

Marsh, H.W. and 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.

Murray, H.G. (1984). Impact of Formative and Summative Evaluation of Teaching in North American Universities. *Assessment and Evaluation in Higher Education* 9: 117-132.

Murray, H.G. (1985). Classroom Teaching Behaviors Related to College Teaching Effectiveness. In J.G. Donald and A. M. Sullivan (Eds.), *Using research to improve university teaching*. San Francisco: Jossey-Bass.

Murray, H.G. (1997). Does Evaluation of Teaching Lead to Improvement of Teaching? *International Journal of Academic Development*, 2: 8-23.

Renaud, R.D. (2002). *The Effect of Higher-order Questions on Critical Thinking Skills*. Unpublished doctoral dissertation. University of Western Ontario, London, Ontario, Canada.

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

Worthington, A.G. and Wong, P.T. (1979). Effects of Earned and Assigned Grades on Student Evaluations of an Instructor. *Journal of Educational Psychology* 71: 764-775.

Slide 1

Rationale of Rating Forms for Student Evaluation of Teaching

In lieu of measuring student learning directly, student rating forms are intended to measure characteristics of teachers and courses that are:

- believed to contribute to student learning
- observable by students
- widely applicable
- under control of instructor

Slide 2

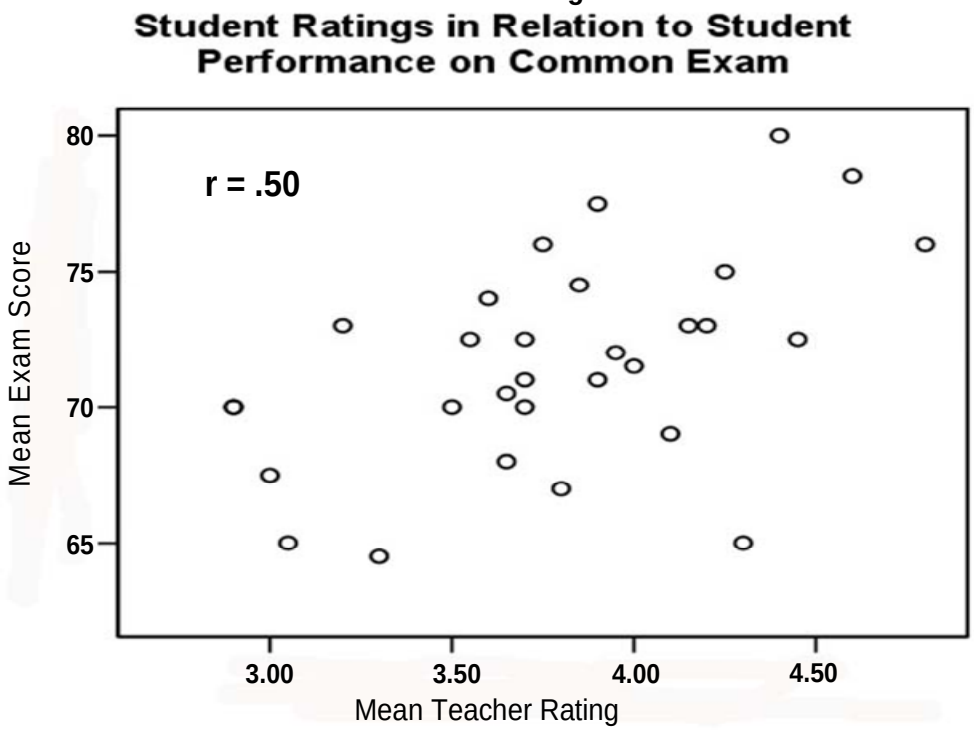
Are Student Ratings Valid?

Predictability of student ratings from specific, low-inference teaching behaviours (Murray, 1985)

Teaching Behaviour	Correlation With Teacher Rating
Clarity	
stresses important points	.47
signals transition to new topic	.31
Expressiveness	
speaks expressively	.63
gestures with hands and arms	.34
Interaction	
addresses students by name	.32
asks questions of class	.37

Slide 3

Are Student Ratings Valid?



Slide 4

Impact of Student Evaluation of Teaching (SET) on Faculty Personnel Decisions

Results obtained from three types of research designs:

- Faculty Opinion Surveys
  average weight placed on SET = 20- 30%
- Field Studies of Actual SPT Decisions
  SET accounts for 10% of variance
- Simulation Studies
  SET has significant but small impact

Slide 5

Does Student Evaluation of Teaching Lead to Improvement of Teaching? (Murray,1997)

Research evidence is available from three types of studies:

- Faculty Opinion Surveys
- Field Experiments
- Longitudinal Comparisons

Slide 6

Does Student Evaluation of Teaching Lead to Improvement of Teaching?

Results of Faculty Opinion Surveys

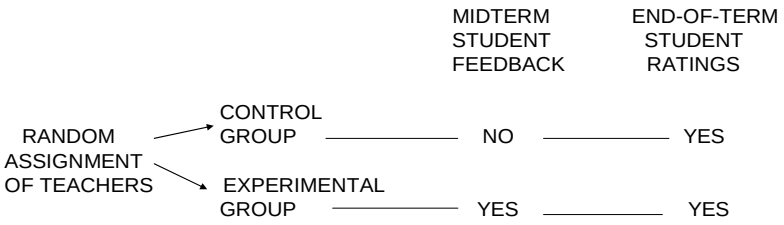
Average percent agreement from eight studies:

Do student evaluations provide useful feedback for improvement of teaching?	73.4%
Have student evaluations led to improved teaching?	68.8%

Slide 7

Does Student Evaluation of Teaching Lead to Improvement of Teaching?

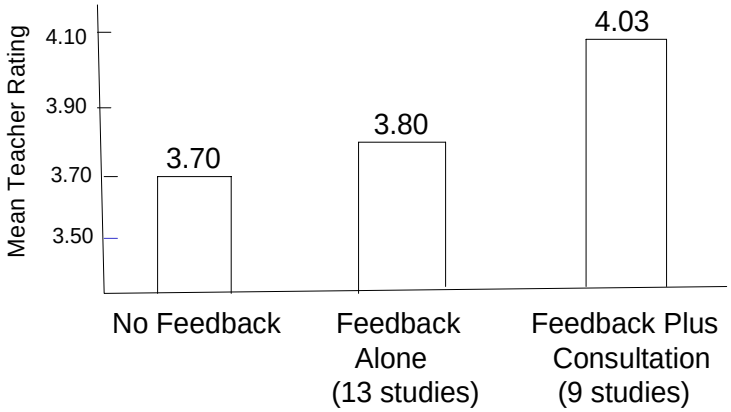
Design of Field Experiments



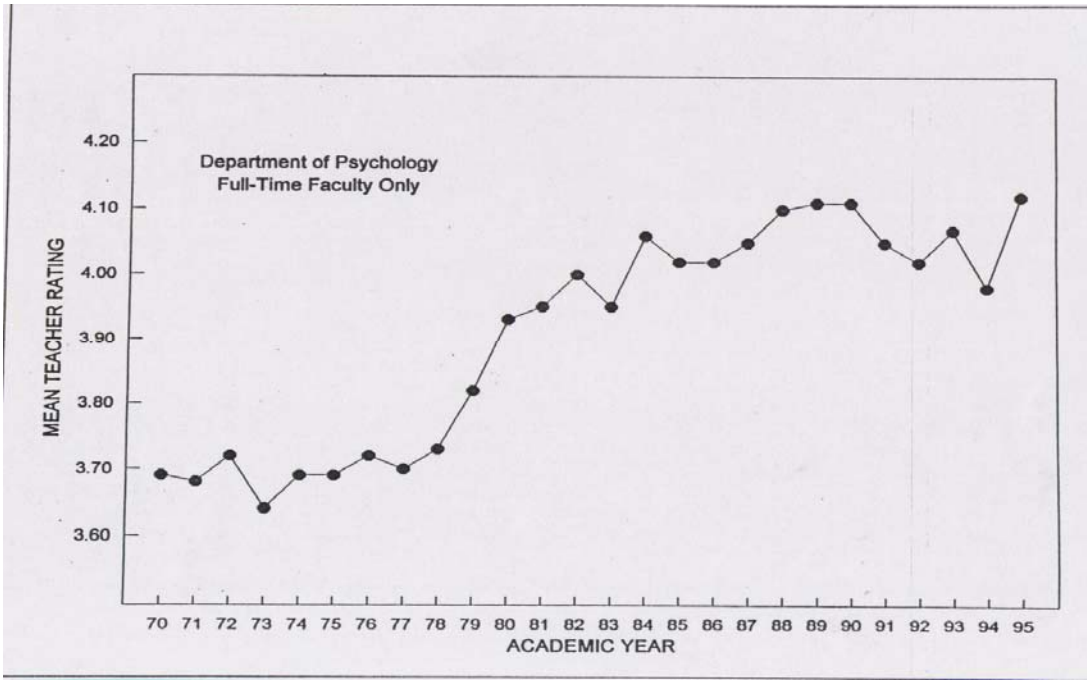
Slide 8

Does Student Evaluation of Teaching Lead to Improvement of Teaching?

Results of Field Experiments (Cohen, 1980)



Slide 9



Slide 10

Impact of Student Evaluation of Teaching on Academic Standards

Do Student Ratings Cause Grade Inflation?

The correlation found between severity of grading and student ratings of teaching could reflect several different causal patterns:

Grades →	Ratings	The grade inflation hypothesis
Teacher Effectiveness →	Grades → Ratings	The teacher effectiveness hypothesis

Slide 11

Some Possible Alternatives or Supplements to Student Evaluation of Teaching

- direct measurement of student learning
- performance indicators
- colleague evaluation of teaching
- NSSE survey