DOCUMENT RESUME

ED 189 920 HE 012 818

AUTHOR Murray, Harry G.
TITLE Student Evaluation of University Teaching: Uses and Abuses. Report No. 5.
INSTITUTION British Columbia Univ., Vancouver. Faculty of Education.
PUB DATE Dec 79
NOTE 50p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Faculty: *Evaluation Criteria: *Evaluation Methods: *Faculty Evaluation: Feedback; Higher Education: Peer Evaluation: Performance Factors: Questionnaires: Reliability: Standards; *Student Evaluation of Teacher Performance; *Student Teacher Relationship: Surveys: Teacher Effectiveness: Validity

ABSTRACT

The percentage of universities using student ratings to evaluate teacher performance is steadily increasing, but this method can be abused if it is the sole criterion of evaluation. Recent surveys of faculty attitudes towards teaching evaluation indicate that most faculty members are in favor of systematic evaluation, prefer student ratings to other methods of evaluation, and believe that student ratings should be used as one of several sources of information in salary, promotion, and tenure decisions. A comprehensive, fair teaching evaluation form to be used for summative purposes should be: easily understood by students; related to student learning: satisfactory in terms of psychometric criteria; under the instructor's personal control: and applicable to many different types or styles of teaching. Student ratings, to be reliable, should be examined for internal consistency, inter-rater reliability, and test-retest reliability. Research on the validity of student ratings is reviewed, and sources of bias are identified. Alternatives to student ratings in the evaluation of teaching might be direct measures of student achievement and colleague evaluations. Considering these alternatives, and certain limitations of student evaluations, student ratings of teaching are sufficiently reliable and valid to be used only as one of several sources of information in administrative decisions relating to faculty salary, tenure, and promotion. References are appended. (DC)

STUDENT EVALUATION OF UNIVERSITY TEACHING:

USES AND ABUSES

Harry G. Murray

University of Western Ontario

Report #5

December, 1979

Colloquium presented at The University of British Columbia, Vancouver, Canada, March 1979, sponsored by Standing Committee on Teaching (SCOT) and Centre for the Improvement of Teaching and Evaluation (CITE) of the Faculty of Education.

Preface

On 19 and 20 March 1979, The Standing Committee on Teaching was pleased to welcome Dr. Harry G. Murray to the U.B.C. campus as visiting consultant/speaker on university teaching. Dr. Murray, from the Department of Psychology, University of Western Ontario, is a leading researcher in the area of evaluation of university teaching.

Dr. Murray's visit continued a series begun in 1978 by the appearance of Dr. Patricia A. Cranton of McGill University who spoke to and consulted with U.B.C. faculty members for two days then. Evaluating teaching for improvement, the area of Dr. Cranton's expertise resulted in a colloquium paper that has been produced as CITE Report Number 1 (1978).

During his stay at U.B.C., Harry Murray conducted a graduate seminar on research in evaluating teaching, held consultations and informal discussions with SCOT/CITE staff, faculty and graduate students, and delivered the colloquium covering his work and that of others on evaluating teaching.

Now in response to a number of requests, we are happy to make available Dr. Murray's colloquium paper on the topic of student evaluation of teaching for personnel decisions. This paper is produced here as CITE Report Number 5.

The Standing Committee on Teaching hopes and intends that distribution of this report through the Centre for the Improvement of Teaching will facilitate discussion and impact of the points Dr. Murray has raised. Recipients of this document, having comments or questions concerning its contents are encouraged to communicate directly with the author at the Department of Psychology, University of Western Ontario, London, Canada, or with the sponsors, SCOT/CITE, Faculty of Education, University of British Columbia, Vancouver, B.C., Canada.

For additional copies of this paper or others in the CITE Report series, past or forthcoming, send requests and $2.00 for each report desired to the Standing Committee on Teaching.

Stephen F. Foster
Vancouver, B.C.

December 1979

STUDENT EVALUATION OF UNIVERSITY TEACHING:

USES AND ABUSES

Harry G. Murray

University of Western Ontario

Evaluating teaching in higher education is a very controversial topic. There are three general methods by which teaching has been evaluated - colleague ratings, student ratings, and direct measurement of student achievement.[1] I am going to discuss only one of these methods today - namely, student ratings. One reason for focusing on student ratings is that this method of evaluation is used more frequently than any other. Recent surveys indicate that approximately 68% of universities in North America have some sort of student evaluation programme, and the percentage of universities using student ratings is steadily increasing (Bejar, 1975). A second reason for focusing on student ratings is that most of the research on teaching evaluation in higher education deals with student ratings, with well over 1,000 articles and books published on this topic. What I will do today is to give an overview of this vast research literature, and at the same time point out some of the potential uses and abuses of student ratings as measures of university teaching.

[1] Self evaluation and data from diverse sources have not received sufficient research and development work to date and are not discussed here.

Purposes of evaluation

Before discussing research on student ratings of teaching it would be a good idea to remind ourselves of the purposes or goals of evaluating university teaching. Four purposes are usually listed. One is called summative evaluation, which means evaluating teaching for the purpose of making personnel decisions concerning salary increments or the granting of tenure or promotion. This is the single most controversial topic in evaluating university teaching, for obvious reasons; and most of my remarks today will be directed to the use of teaching evaluations for summative purposes.

For the past ten years, my own university, the University of Western Ontario has regularly used student ratings for promotion and tenure purposes. Many faculty members at Western are not aware that teaching evaluations could be used for any other purpose than summative evaluation. At some universities or colleges it is just the opposite - that is, faculty members support the use of student evaluations for a second purpose, that is formative evaluation, or improving teaching, but are reticient about using student ratings for deciding tenure or promotion.

The third purpose of evaluating teaching is to provide information to be used by students in selecting courses. At some universities teacher evaluations are published annually in booklets or "anti-calendars". We have a publication of this type at Western Ontario, but faculty members have the option of whether or not to allow publication of their ratings.

Most allow their ratings to be published annually in the anti-calendar. In the Faculty of Social Sciences last year only two or three out of about 250 faculty members refused permission for their ratings to be published.

The fourth and final purpose of evaluating teaching is to provide a quantitative measure for research projects on university teaching. This purpose is very important to me because I have kept myself off the streets for the past ten years by doing research on university teaching!

One very important point in relation to the various purposes of teaching evaluation is that, as a general rule of thumb, it is a mistake to try to use a single teacher rating instrument for more than one of the different purposes we have discussed (Glasman and Gmelch, 1976). Some universities, including my own, try to do just that. This practice is bound to cause problems because different types of evaluative data are needed for different purposes. For example, the type of questionnaire items most suitable for promotion and tenure purposes - that is global, standardized items which permit comparison of instructors teaching different types of courses - would be next to useless for purposes of formative evaluation or improving teaching. Being told that you are unsatisfactory in terms of "clarity" or "rapport" or "overall effectiveness" does not help much in improving teaching, because there is no indication as to why you are perceived this way or what you might do to improve. To be useful for formative purposes, student evaluations should be specific and individualized rather than global and standardized, and should include specific suggestions for improvement. And to be useful for purposes of student course selection, evaluations probably should focus as much on characteristics of the course as on skills of the teacher. It is extremely unlikely that one rating form can satisfy all of these requirements simultaneously.

Faculty attitudes toward teaching evaluation

Recent surveys of faculty attitudes towards teaching evaluation indicate that most faculty members: (1) are in favour of systematic evaluation of teaching, (2) prefer student ratings to other methods of evaluation, and (3) believe that student ratings should be used as one of several sources of information in salary, promotion, and tenure decisions (e.g., Goldstein and Anderson, 1977). A study done at various colleges and universities in California by Rich (1976) showed that across all institutions, 75% of the faculty members favoured the use of student evaluations in promotion and tenure decisions. Furthermore, and contrary to expectation, it was found that attitudes toward student evaluation were more favourable at research-oriented universities than at institutions giving primary emphasis to teaching. The reason for this is not clear. My hypothesis is that student ratings are less threatening to faculty members at research-oriented universities because these people have something else to fall back on besides student ratings of teaching. If the ratings are poor, at least one can be active in research. A teacher need not put "all his eggs in one basket". On the other hand, student ratings are likely to be perceived as more threatening at a college that specializes only in teaching, because there one must rise or fall on the basis of student ratings alone. In any case, the overall result of Rich's study, like that of similar studies, was that faculty members were surprisingly favourable toward student ratings as a measure of their teaching competence.

Rationale of teacher evaluation instruments

I would like to say something about the guidelines that should be taken into account in the development of a teacher evaluation instrument. The typical questionnaire or rating form for student evaluation of teaching focuses on factors such as clarity, preparation, enthusiasm, rapport, and quality of feedback, which are (1) observable by students, and (2) believed to be causally related to student learning. In other words, the rationale of a student rating form is to assess the process of teaching, and from this to infer the probable outcome of teaching. We would all agree that the best teacher is the one who produces the highest level of learning in students. Unfortunately, it is very difficult to assess those components of student learning which are uniquely attributable to skills of teachers. Thus we do the next best thing in asking students to rate characteristics of their teachers which one would logically expect to be determinants of student learning.

In selecting the specific items to be included in a student rating form, we would consider normal psychometric criteria such as clarity of wording, discriminatory power, reliability, and statistical independence or nonredundancy. But in addition to these criteria, there are some special precautions that must be taken into account in the development of a rating form to be used for promotion and tenure purposes. The most important of these, and one which is all too frequently ignored, is that the rating items must refer to aspects of teaching which are under the instructor's direct control, and for which the instructor can be held personally accountable. Obviously it makes no sense to judge an instructor for promotion or tenure on the basis of factors such as the availability or quality of the textbooks, which are largely outside the instructor's control. Student ratings of the quality of the textbook might be quite useful for purposes of improving the course or for purposes of informing students vis à vis course selection, but are totally unsuitable for use in promotion and tenure decisions because there simply may be no good textbook available in some areas. Text availability may depend upon transport, labour or weather problems even when the instructor ordered it in a timely fashion. Another item which falls in this category is the instructor's availability for consultation with students. It is not fair to compare someone teaching 8 students to someone teaching 400 students on such an item. Another problematic area is the interest value of the course, or "relevance" of the course. Clearly students are more willing to assign high interest or relevancy ratings to certain types of courses, such as abnormal psychology, than to other types of courses, such as psychological statistics. A statistics teacher has a real disadvantage on an item like this, if it is used in promotion and tenure decisions. According to my observations, students will say the statistics teacher did a good job of presenting the material, but most students are unwilling to say that the material itself was interesting, or had an impact on their life, or was relevant. Things that are out of the instructor's control, however interesting and informative for other purposes, should not be used for promotion and tenure purposes.

A second requirement of a teacher rating form intended for promotion and tenure purposes is that the items be applicable to a wide range of different types and levels of teaching. Only through the use of standardized, widely applicable items of this sort is it possible to have a common basis for summative comparison of all of the members of a faculty or department.

In summary, we have seen that the items included in a teacher evaluation form to be used for summative purposes must be: (1) observable by students, (2) related to student learning, (3) satisfactory in terms of psychometric criteria, (4) under the instructor's personal control, and (5) applicable to many different types or styles of teaching. Clearly a great deal of thought and effort must go into the development of a teacher evaluation instrument that satisfies all these criteria.

Another issue that comes up in connection with the development of student rating forms is the question of multiple-choice versus open-ended items. In some disciplines, particularly in the arts and humanities, the preference is strongly on the side of open-ended prose evaluations - with multiple-choice items viewed as the root of all evil in the world, something invented by psychologists and to be avoided by all others! I would not want to dissuade anyone from what they believe in their hearts to be the only proper way of evaluating something as complex and individualistic as university teaching. However, we must remember that after reading all of those hundreds of pages of prose commentary, we will eventually get to the point of coming up with a summary rating or summary evaluation that will be used by the promotion and tenure committee. And I would wager my last dollar that the results obtained in this way would correlate very closely with comparable summary evaluations derived from multiple-choice items. In other words, I am saying that multiple-choice and open-ended questions will give exactly the same results ultimately, but multiple-choice questions will do so with considerably less time and effort. This is not to say that open-ended questions are useless, but rather that their most effective use is in a supplementary role, as a source of diagnostic feedback for teaching improvement.

Regardless of how much work goes into development of a questionnaire for student evaluation of teaching, such an instrument will inevitably have a number of serious limitations. For one thing, a student rating form can measure only those things that happen in the classroom, and are therefore observable by students. Aspects of university teaching which go on behind the scenes, so to speak, like writing textbooks, individual consultations with students, preparing laboratory materials, curriculum development, course design, and thesis supervision, are all part of teaching but are not measured by student rating forms which focus on classroom instruction. Another restriction is that student ratings can only measure the "delivery" aspect of instruction, and not the "content" aspect, because students are not in a good position to judge content. If students were in a good position to judge content, they would not be students! Student ratings can tell us whether the material that is taught is taught effectively, whether it comes across well, but they can not tell us whether the material itself is accurate, or up to date, or academically significant.

Judgements of the latter sort should be left to colleagues or subject matter experts.

Reliability of student ratings

Let us next consider the reliability of student ratings. By "reliability" I mean the extent to which ratings are consistent or dependable for a given teacher. If a teacher receives a certain rating in a course this year, how likely is it that he or she will receive a similar rating in another course next year? Reliability is not the same as "validity", which refers to whether or not student ratings measure that which they are supposed to measure. In the case of reliability, we are asking whether student ratings measure something consistently, regardless of whether this something is that which is intended. Researchers have investigated the reliability of student ratings by comparing mean ratings of the same teachers across: (1) sets of items, (2) raters, (3) time periods, and (4) different courses. Table 1 provides a summary of results obtained in these four types of comparisons, which I will discuss in turn (cf., Doyle, 1975; Murray, 1979).

One way of looking at the reliability of student ratings is in terms of consistency of ratings across items on a questionnaire. This is called inter-item reliability or internal consistency reliability. Mean student ratings for a given teacher are very reliable in this sense. If you have a set of items, say six items, that are supposed to measure the same aspect of teaching, and you compute pairwise correlation coefficients between items, the average correlation coefficient will be very high, probably around .85 to .95. In other words, if students rate the teacher highly on one item they will normally rate him or her equally highly on other items intended to measure the same behaviour or trait.

Another type of reliability is inter-rater reliability. Here the question is "Do students agree with one another in the ratings they give?". Usually inter-rater reliability is computed by randomly dividing students in a class into two equal groups, then computing two separate mean ratings, one for even-numbered and one for odd-numbered students and correlating "odd" with "even" mean ratings across teachers. Normally when this is done the inter-rater reliability coefficient is quite high, say .80 or higher. Centra's (1973a) work indicates that as long as there are a minimum of 15 students in a class the reliability of the mean teacher rating will be quite high. If there are less than fifteen students in a class the reliability drops off considerably. It is probably unwise to use student ratings based on less than ten students, as a rule of thumb.

The third approach to reliability, usually called test-retest reliability, is more stringent. In this case reliability is measured by correlating teacher ratings at two points in time for the same course or same type of course. In some studies, the two points in time fall in the same course and same year. The reliability coefficient in this case is very high, usually .70 to .80. If the teacher receives a high or low rating at the middle of the term, he is likely to receive a similar rating at the end of the term. In one study using this paradigm, Kohlan (1973) found that ratings obtained after only two class meetings correlated quite highly ($\underline{r}$ = .70) with ratings obtained at the end of the course. It is

TABLE 1

SUMMARY OF RESEARCH FINDINGS ON THE RELIABILITY OF STUDENT RATINGS

Reliability Test	Reliability Coefficient
Reliability of ratings across items (internal consistency)	.85 to .95
Reliability of ratings across raters (inter-rater reliability)	.80 to .90
Reliability of ratings across time periods for same type of course (test-retest reliability)	.70 to .80
Reliability of ratings across different types of courses taught in same or different years	.30 to .40

unclear exactly what this result means. Kohlan claims that students form impressions early in the course on the basis of very limited evidence, and are then unwilling to change their minds. Another interpretation is that good teachers are successful in the very first class, and continue to be successful thereafter. According to this view, the high correlation found by Kohlan reflects actual behaviour of teachers rather than student "bias".

We have seen, then, that within a given course and a given year ratings are reliable from one point in time to another. Another way of assessing test-retest reliability is to compare ratings in two or more successive years for the same course or same type of course. This is done by locating a sample of teachers who have taught the same course in two successive years and computing the correlation coefficient between mean ratings in Year 1 and mean ratings in Year 2. The reliability of ratings continues to be quite high, around .75, in most studies done with this paradigm. Thus there is still a high level of consistency in student ratings as long as we stick to comparisons across time for a given course or type of course.

The fourth and final way of investigating the reliability of student ratings is to compare ratings of the same teachers across different courses or different types of courses. (I considered but eventually rejected the idea of labelling this type of comparison "inter-course reliability"). Studies using this paradigm (e.g. Hogan, 1973) have reported surprisingly low reliability coefficients, usually in the .30 to .40 range. In other words, student ratings of a given teacher are not particularly stable across different types of courses. If an instructor teaches a seminar in abnormal psychology to fourth year students and a large introductory psychology lecture class to freshmen in the same year, the ratings he receives in these two courses may be quite different. If teacher ratings are compared across different courses taught in different years, the reliability coefficient is likely to be lower still. It would appear, then, that it is somewhat oversimplified to conclude that student ratings of teaching are "reliable". Student ratings are reliable as long as we are talking about ratings across time for a given type of course. They are considerably less reliable across different types of courses taught in the same year, or in different years. This means that teaching ability is not the highly generalized trait that people often think it is. If an instructor is good at one type of course it is not necessarily true that he will be good in other types of courses. Teaching tends to be more course specific than is generally believed. In fact analyzing data for individual instructors one can see that there are some teachers who consistently get low ratings in introductory courses and high ratings in fourth year honours courses, and other teachers who do just the opposite, that is, receive high ratings in large, lower-level classes and low ratings in senior seminars. In other words, some teachers seem best suited to teaching introductory lecture courses, whereas others are suited for small-group teaching. This is not to deny that there may exist some master teachers who excel in many or all types of teaching, but this is the exception rather than the rule.

One important implication of the low correlation of ratings across courses is that using student ratings for promotion and tenure purposes requires a good sample of ratings from different years and different types of courses. Hogan (1973) says that student ratings must be available for at least six different courses in order to provide a sufficiently reliable measure for something as important as a tenure or promotion decision.

A second implication of the low correlation of ratings across courses is that instead of assuming that instructors are equally suited for all types of courses, department chairmen might try to maximize departmental teaching effectiveness by assigning instructors to those courses in which they have received their highest ratings in the past. Of course, such a plan would be put into effect only after a faculty member has tried his hand at different types of courses, and perhaps has already been granted tenure. Although this plan would probably result in improved teaching, I must admit that it might be difficult to persuade someone who is outstanding in teaching large freshmen courses that it is in everyone's best interests for him or her to continue doing this sort of thing all the way to age 65!

A third implication of the lack of consistency of ratings across courses is that the use of teaching ratings by students in selecting courses is somewhat suspect. If a student looks at an anti-calendar and finds that Professor Jones received a certain rating for course X, and the student is considering taking course Y from the same professor, the data suggest that student ratings may not be very useful for this purpose. The predictability of Professor Jones' ratings in course Y, as indicated by course X, is not likely to be high. Thus, the student is probably not improving much upon random guessing in using ratings to select courses in this way. From the student's point of view, using teacher ratings from course X to decide whether to take course X from the same teacher is a more sensible strategy.

Validity of student ratings

The next question is the most important of all - the question of validity of student ratings. Do student ratings measure that which they are intended to measure? We should first remind ourselves of what it is that student ratings are intended to measure. As noted previously, the typical student rating form is designed to assess behaviours or characteristics of the teacher that are believed to be related to student learning. Thus there really are two questions concerning the validity of student ratings. The first question is "Does the rating form really measure the traits we think it measures, as opposed to something else?". The second question is "Are these traits related to student learning?". It is possible for a student rating form to be valid in the first sense, but not valid in the second sense. Table 2 summarizes the three different approaches that have been used in studying the validity of student ratings.

Sources of bias in student ratings. The first approach has been to see whether student ratings are correlated with extraneous factors such as class size, severity of grading, time of class meeting, and whether the course is required or optional. The logic of this approach is that if student ratings validly measure teacher behaviours such as clarity, enthusiasm, and preparation, then the results - the ratings - should be relatively unaffected by extraneous factors. If the teacher explains clearly it should not matter whether there are 300 students or 20 students in the class, or whether high or low grades are awarded, the ratings for explaining clearly should be high. Thus it is expected that there should be a negligibly low correlation between ratings and extraneous factors if student ratings are "valid". Research on this issue indicates that, contrary to expectation, student ratings are in fact significantly affected by a number of extraneous factors. Some authors have suggested that the effect of extraneous variables is minor enough to be ignored, but this is an over-simplification. Clearly there are some valid, genuine biasing effects of extraneous variables that are not easily ignored. For instance, there is a negative correlation of approximately -.20 between class size and ratings given to the teacher. This correlation is as high as -.35 in some studies, but averages around -.20, which means that there is a small but significant tendency for bigger classes to give lower ratings than small classes (e.g., Crittenden, Norr & LeBailly, 1975). Correlations of opposite sign but similar magnitude, around .25 to .35, have been reported between teacher ratings and grades assigned to students (Feldman, 1976). In other words, teachers who assign low grades tend to receive lower ratings from students than teachers who give high grades. One explanation of this relationship is that students "reward" easy grading teachers with high ratings and "punish" strict graders with low ratings. Some observers claim that this contingency will lead to a continuous cycle of grade inflation and abandonment of academic standards. It is not clear to what extent this is true, however, because there is some ambiguity about what the grades-ratings correlation implies. It is conceivable that this correlation simply reflects a tendency of better teachers to produce higher levels of learning in their students. Since

TABLE 2

RESEARCH ON THE VALIDITY OF STUDENT RATINGS OF TEACHING

Sources of bias in student ratings
1. class size
2. severity of grading
3. time of class meetings
4. percent attendance for ratings
5. teacher reputation
6. the "Dr. Fox effect"

Student ratings and ratings of others
1. alumni
2. classroom observers
3. colleagues

Student ratings and student achievement
1. Rodin and Rodin (1972)
2. Sullivan and Skanes (1974)
3. other studies

better teachers produce higher levels of learning, their students would legitimately be expected to receive higher grades. For the present it is not possible to disentangle these factors and to ascertain the extent to which the correlation reflects "bias" as opposed to a genuine impact of the teacher upon student performance.

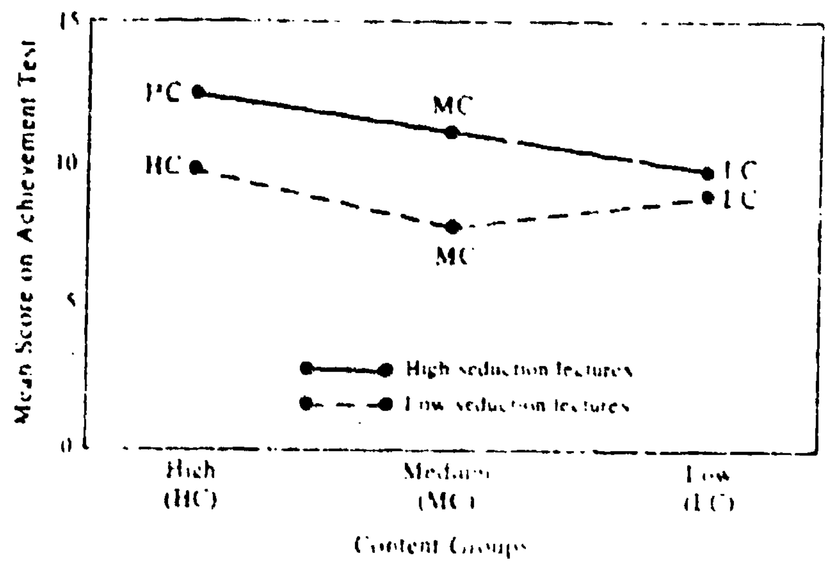
In the case of the time classes meet, results are mixed. One might expect that classes meeting at 8:30 a.m. would get lower ratings, but that does not seem to be substantiated in the literature. If anything, classes meeting at mid-day get lower ratings, perhaps because education is interfering with lunch! However, results are inconsistent here, and the most justifiable conclusion is that time of class is not a source of bias in student ratings.

Another bias variable that researchers have looked at is percentage attendance on the day when ratings are taken. The results show that if the level of attendance is either very low or very high, then ratings are high. If attendance is in the middle, in the 40-60% range, then ratings are lower (Centra and Creech, 1976). I have a notion that in the case of classes with low attendance, the students who are there when ratings are taken are "chronic attenders" who are not driven away by even the poorest of teachers and who are predisposed to give high ratings. If a poor teacher is lucky enough to have mostly chronic attenders on the day of reckoning, then the ratings will be spuriously high! Whatever the reason, percentage attendance appears to be yet another source of extraneous variance in student ratings.

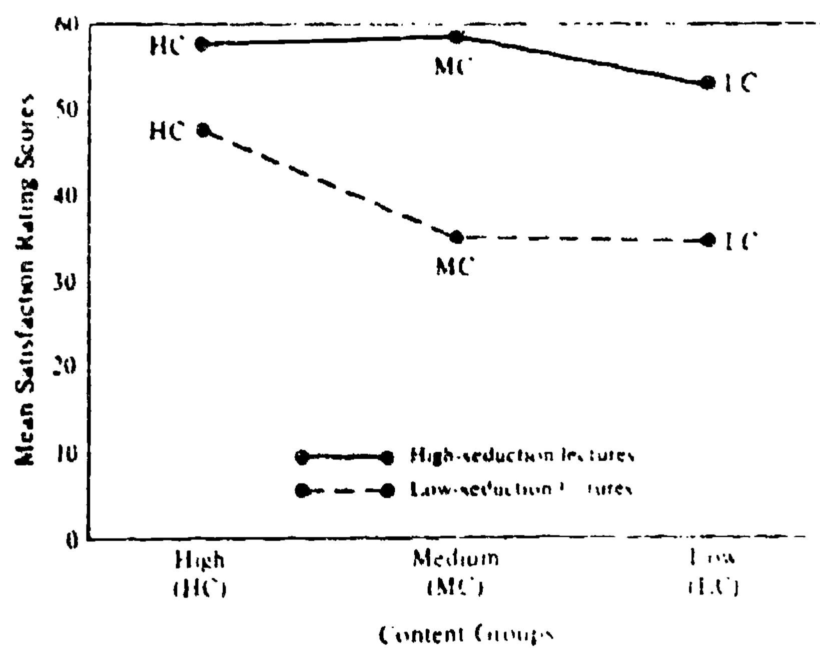
Yet another variable that has been studied as a potential source of bias is teacher reputatio... It is claimed that a faculty member who has a good reputation as a teacher will receive higher ratings solely on the basis of his or her reputation. Supposedly such individuals are positively perceived largely on the basis of their reputations. I do not put much credence in this. If I was in a class and was told that the instructor was a good teacher - in other words, he or she has received high teacher ratings - and I was subjected to dull, boring, muddled lectures for eight solid months, I do not believe that I would give that teacher a high rating at the end of the year. It is claimed that this happens, however.

At the University of Manitoba, Leventhal, Abrami, and Perry (1976) have investigated a more sophisticated (and more plausible) type of teacher reputation effect. They divided students into categories according to why they registered in a particular section of a multiple-section course. One of the most common reasons for selecting a particular class section was teacher reputation - about 30% of students said that they selected on this basis. The most important point is that students who took the course because of teacher reputation tended to rate the teacher more positively than students who took the course for other reasons. Now a teacher who has received high ratings in the past and has a good reputation will supposedly have more of this type of student in his or her class. It is claimed that such a teacher has a distinct advantage in the sense that a large proportion of his students have a predisposition to give high ratings. Although such a biasing effect is theoretically plausible, it is difficult to assess the extent to which it operates in actual practice. For example, the effect would be nullified under conditions where students are assigned to class sections by lot or solely on the basis of timetable constraints.

A final source of bias in student ratings which has received considerable attention in the literature is the so-called "Dr. Fox effect". Teachers with little new to say but say it in a very enthusiastic expressive way supposedly are able to "seduce" students into giving them high ratings. This phenomenon has been investigated largely under laboratory conditions, with video-taped lectures (e.g. Ware and Williams, 1975). A professional actor (introduced as Dr. Fox) is hired to produce a series of video-taped lectures which vary both in level of content and in expressiveness of presentation. In order to create high, medium, and low levels of content coverage, the same lecture is re-taped with more and more substantive points omitted. High and low levels of expressiveness are created through selective use of voice inflection, humour, movement and gesture, and enthusiasm. Separate groups of students view video-taped lectures embodying six different combinations of content and enthusiasm - three levels of content coverage and two levels of enthusiasm. The results of one of the "Dr. Fox" studies are summarized in Figure 1. First, it is found that students learn more (as indicated by an immediate recall test) with high content coverage than with low content coverage, and learn more when the lecture is presented enthusiastically than when presented without enthusiasm. Second, it is found that, on the average, student ratings of lecture quality are higher for high content than for low content lectures, and higher for enthusiastic than for nonenthusiastic lectures. So far we have not touched on the Dr. Fox <u>effect</u>. The Dr. Fox effect refers to the fact that under the high enthusiasm condition, student ratings are not sensitive to content differences. That is, students rate an enthusiastic lecture equally highly regardless of its level of content coverage. It is claimed on the basis of this result that students can be seduced into giving high ratings by an enthusiastic but uninformed teacher, and thus student ratings are "invalid". My own view is that although the Dr. Fox effect is interesting and has attracted much attention, it has very little or no relevance to real teaching in real classrooms. For one thing, I doubt that there are many "Dr. Fox" teachers on the faculties of reputable universities. In the first Dr. Fox experiment, an actor was hired to give a talk on "Mathematical Game Theory Applied to Physical Education" to a group of educators, psychologists, and social workers. He certainly knew nothing about this subject (I'm sure no one else knows anything about it either), but he presented his talk very enthusiastically and was given high ratings. Not only is it implausible that there are many faculty members like this in reputable universities, but even if there were, it seems unlikely that students would be fooled by such a teacher for a whole year, especially if their performance on examinations depended on the information the person conveyed. By checking the teacher's knowledge of the subject matter against the textbook students would discover he did not know the material, and would



Mean scores on the achievement test of students who experienced lectures high and low in "seduction" (that is, enthusiasm) at three levels of content



Mean scores on the satisfaction rating scales of students who experienced lectures high and low in "seduction" (that is, enthusiasm) at three levels of content

Figure 1. Results on the Ware and Williams (1975) study of the Dr. Fox effect. Top panel shows student achievement, bottom panel shows student ratings.

react negatively. Thus my own view is that the Dr. Fox effect cannot be generalized from the laboratory to a real teaching situation. But it is still instructive and reminds us again that since students can probably only rate delivery well, and not content, student ratings must be supplemented by colleague evaluations of course content. Another lesson we can learn from the Dr. Fox studies is to be more enthusiastic in our teaching. As may be seen in Figure 1, one of the major findings of the Ware and Williams study was that both student learning and student ratings were higher for enthusiastic than for nonenthusiastic lectures. The implication is that university lecturers can improve both their teaching evaluations and their impact upon student learning (who could ask for anything more?) simply by lecturing with more enthusiasm or expressiveness. This aspect of the Dr. Fox studies is perhaps the most important and most useful of all, but has been totally ignored by the original investigators in favour of more rhetorical issues.

I would like to summarize the results on bias effects in student ratings. I have spent a lot of time discussing this topic, partly because more research has been done on bias effects than on any other aspect of student ratings. This research has demonstrated that student ratings are significantly affected by a number of extraneous factors, including class size, severity of grading, teacher reputation, and teacher "seductiveness". These sources of bias cannot be ignored by anyone responsible for interpreting student ratings in administrative personnel decisions. On the other hand, we must keep in mind that correlations found between ratings and bias factors, although statistically significant, have typically been rather small in absolute terms. Studies by Brown (1976) and others using multiple regression techniques indicate that several bias factors in combination account for no more than 15% of the variance in teacher ratings. Furthermore, in the case of some bias factors (e.g. severity of grading) it is possible to attribute results obtained to mechanisms other than "bias". And in the case of other bias factors (e.g., teacher seductiveness), there is some doubt as to whether results demonstrated in the laboratory will generalize to real-world teaching conditions. Even when effects of bias factors can be demonstrated unambiguously, it is sometimes possible to minimize or eliminate these effects through the use of statistical adjustments or separate norm groups for different types of courses (e.g. required vs. optional, large vs. small). For these reasons, it can be argued that the contribution of bias factors to student ratings is neither large enough nor clearcut enough to single-handedly invalidate student ratings as a measure of teaching effectiveness. Nevertheless, anyone facing the responsibility of using teacher evaluation data in promotion and tenure decisions must be fully aware of potential sources of bias in student ratings, and must take full acount of these biases in arriving at any decision.

Student ratings and ratings of other observers. A second way of checking the validity of student ratings is to correlate them with ratings of the same teachers made by other observers, such as colleagues, alumni, or trained judges. This approach is analogous to checking the accuracy of an unknown measuring instrument by comparing it to a trusted or pre-established instrument. Three different studies, including one by Centra (1974), have investigated the relationship between teacher ratings given by current students and those given by alumni of several years standing. In general, these studies show that current students and alumni are in fairly close agreement as to who is a good teacher and who is a poor teacher (inter-rater correlations range from .60 to .75). This finding contradicts the common idea that students are immature and do not have the long range perspective to judge teaching properly. Of course, one might wonder how mature some alumni are after observing their performance on homecoming weekend!

Other studies (e.g., Murray, 1972, 1977) have compared student ratings to those of trained observers who are paid to visit class sessions and provide a "neutral" or "objective" view of teaching. Typically the observers record specific behaviours of the instructor as well as providing global ratings of teaching effectiveness. The general findings of these studies is that student ratings agree very closely with ratings of outside observers. In my own research on this question, I have found a correlation of .92 between student and outside observer ratings of psychology lectures (Murray, 1972), and a multiple correlation of .81 between student ratings of overall teaching effectiveness and outside observers' reports of 10 specific classroom teaching behaviours (Murray, 1977). These results suggest that student ratings are determined more by actual behaviours of the instructor than by extraneous bias factors. In other words, student ratings seem to contain a much larger proportion of "true variance" than of "error variance". Certainly, to send someone who is (1) trained and (2) unbiased into the classroom to observe the teacher, and to find that this person gives you just about the same results as students do, suggests that student ratings do indeed have some degree of validity.

Research on the relationship between student and colleague ratings of teaching (e.g., Doyle and Crichton, 1978) has yielded correlations ranging from very low to very high, with a median value of approximately .50. An obvious reason for the lower correlations found in these studies is that colleagues have little or no opportunity to observe classroom teaching, and thus base their ratings on factors other than those used by students (e.g., scholarly reputation). On the other hand, it can be argued that the high student-colleague correlations obtained in some studies are due to the fact that colleague ratings were influenced by direct or hearsay knowledge of prior student ratings. Because of methodological problems such as these not much can be concluded one way or the other from research on the relationship between student and colleague ratings.

Student ratings and student achievement. The third way that the student rating validity has been studied, and probably the most important of all, has been to determine the correlation between student ratings and objective measures of the teacher that are related to student learning. There should be a significant positive correlation between student ratings and how well the students perform on an achievement test. To put it more simply, teachers who receive high ratings should produce higher levels of learning in their students. The way this question is typically studied is in a multiple-section course where there is a common textbook, common curriculum, and common final examination at the end of the year. With say 15 to 20 class sections it is possible to determine the correlation between the mean rating assigned to the teacher and the mean performance of the teacher's students on the common final examination. In the 30 or so studies that have used this paradigm, a range of different results have been reported - all the way from a -.75 correlation to a +.70 correlation between teacher ratings and student achievement. Figure 2 provides a summary of these studies. It may be noted that although results have ranged widely, most studies have found a moderate positive correlation between student ratings and student achievement, and only one study has found a significant negative correlation. That was the Rodin and Rodin study published in Science in 1972. This study has been severely criticized on methodological grounds, but because it claims to show that student learning is poorest for highest rated teachers, it has attracted considerable attention. In fact, when the study first appeared, an enterprising chap in the Department of Mathematics at the University of Western Ontario went to his department chairman and said that since he had received the lowest rating in the department, therefore he was the best teacher in the department, and should be promoted to full professor immediately! As far as I know he was not promoted. But this illustrates the point that many faculty members are familiar with only one of the hundreds of studies done on student ratings, and that one study is the Rodin and Rodin study. All other studies of the relationship between student ratings and student achievement fail to agree with the Rodin and Rodin study, most showing weak to moderate positive correlations. Of the remaining 29 studies, 4 indicate zero or non-significant correlations between teacher ratings and amount learned by students, and the rest indicate positive correlations ranging from .20 to .70, with a median value of about .40 to .50.

One of the best designed studies in this area was done by Sullivan and Skanes (1974) at Memorial University of Newfoundland (that's on the other coast, in case you have forgotten). One of the unique features of this study was that students were randomly assigned to classes, thus ensuring initial equivalence of class sections in terms of student ability. Another unique feature was that a very large sample of class sections, 130 to be specific, was available for study. Consistent with most other studies, Sullivan and Skanes found a correlation of .40 across sections between mean instructor rating and mean final examination performance. Another interesting result was that the ratings-achievement correlation was higher for experienced, full-time faculty members than for inexperienced, part-time teaching assistants. This result might help to explain why Rodin and Rodin, who used inexperienced teaching assistants as "subjects", found a negative correlation between ratings and achievement.

To summarize the results of these studies, the conclusion seems to be that under most conditions there is a low to moderate positive correlation between amount learned by students and the ratings that students give to

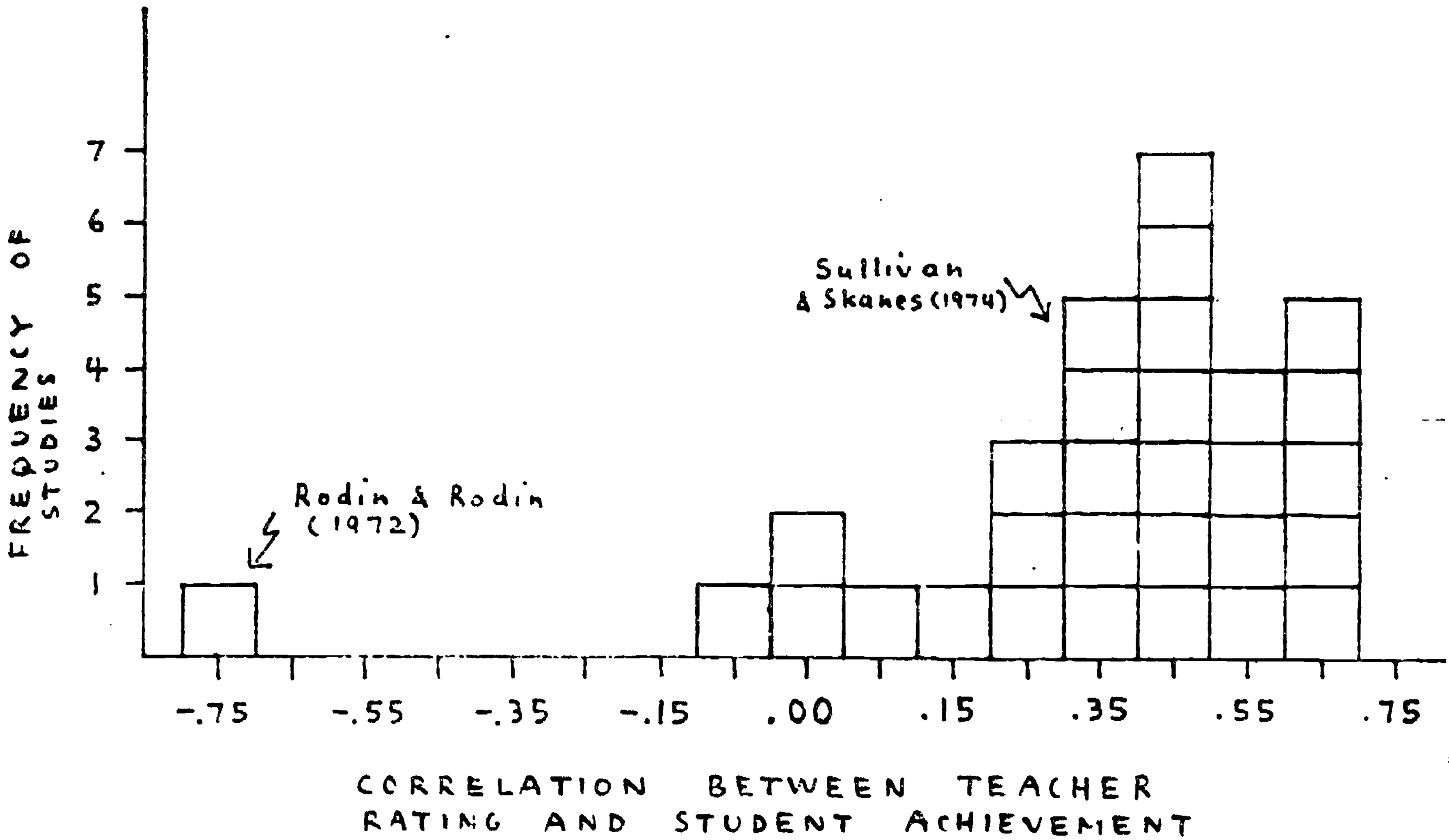


Figure 2. Frequency distribution of teacher rating-student achievement correlations reported in 30 different studies. For studies reporting correlations separately for different rating scales, the median correlation is used here.

their teachers. In other words, teachers receiving high ratings from students tend to produce higher student achievement levels than teachers receiving lower ratings. This relationship between student ratings and student achievement is probably the single most important piece of evidence supporting the validity of student ratings. It must be kept in mind, however, that this is a statistical relationship which is true only in terms of averages for probabilities. Such a relationship leaves plenty of room for exceptions in individual cases.

Putting all of these results together, including reliability of ratings, correlation with bias factors, correlation with other observers, and correlation with student learning, I would offer the following four statements as a fair summary of what we know about the reliability and validity of student ratings of college teaching:

1. Student ratings of a given instructor are highly consistent across different rating forms, subsets of raters, and time periods (e.g. successive years); and are moderately consistent across different types of courses. In other words, student ratings show acceptable levels of reliability or generalizability.

2. Student ratings of classroom teaching ability correlate moderately to highly with comparable ratings made by peers, alumni, and paid classroom observers, indicating that student perceptions of good and bad teaching agree closely with those of more expert or more neutral observers.

3. Student ratings correlate significantly with extraneous bias factors such as class size, severity of grading, and teacher reputation. However, these correlations tend to be small in absolute terms, and in most cases sources of bias can be removed through statistical adjustments or the use of separate norm groups.

4. Although results are inconsistent, the weight of evidence indicates at least a moderate positive relationship between student ratings and objective measures of student achievement. In other words, teachers who receive favourable ratings from students, do in fact promote higher levels of student learning, on the average, than teachers who receive less favourable ratings.

Use of student ratings in administrative personnel decisions

On the basis of the data summarized above, I would argue that student ratings of teaching are sufficiently reliable and valid to be used as one of several sources of information in administrative decisions relating to faculty salary, tenure, and promotion. As I stated earlier, my own university (Western Ontario) has used student ratings for this purpose for nearly ten years, and despite periodic grumblings the program there seems to have been a success. One advantage of using teaching evaluation for promotion and tenure purposes is that this practice communicates to the public that the university takes teaching seriously and holds itself accountable for the quality of its teaching. A second advantage is that faculty members are provided with a tangible incentive for devoting the time and effort that are needed to bring about significant improvement in their teaching. It is unreasonable to expect faculty members to show strong interest in teaching improvement if their efforts in this area go unrecognized in the faculty advancement system.

The research we have just reviewed on the reliability and validity of student ratings can be used as a basis for improving the quality of student rating forms designed for salary, promotion, and tenure purposes.

I noted earlier that the items in such a rating form must have wide applicability and must refer to aspects of teaching that are under the instructor's direct control. Assuming a pool of items satisfying these criteria, the most defensible items for use in salary, promotion and tenure decisions are those which show: (1) the highest levels of inter-rater and test-retest reliability, (2) the lowest correlations with extraneous variables such as class size and severity of grading, and (3) the highest correlations with objective indicators of student learning. Research evidence (e.g., Centra, 1977, Frey, 1978; Seiler, Weybright and Stang, 1977) indicates that for all three of the above criteria, items assessing the instructor's overall teaching effectiveness (e.g., "How would you rate the general teaching ability of this instructor?") are superior to items assessing more specific instructional characteristics; and among items in the latter category, those assessing the instructor's expositional skills (e.g., clarity, speaking ability, use of examples, organization, enthusiasm) are superior to those assessing instructor rapport, student-teacher interaction, and quality of feedback. Thus the best items for use in promotion and tenure decisions would appear to be overall effectiveness items, the next best are those pertaining to expositional skills, and the least defensible items are those relating to rapport, student involvement, and feedback.

Although student ratings may be sufficiently reliable and valid to justify their use in administrative personnel decisions, we must keep in mind that these ratings have several obvious limitations, and therefore should never be treated as a foolproof measure of all aspects of university teaching. It might be useful at this point to summarize some of the limitations of student ratings, most of which have been mentioned previously. The most obvious limitation of student ratings is that they apply only to those aspects of teaching which occur in the classroom. The typical student rating form provides no information abou non-classroom teaching activities such as course design, preparation of educational materials, or individual tutoring of students. It is entirely possible that a person could get low ratings from students but be excellent in non-classroom teaching activities, or vice versa. This would not be clear if teaching were measured solely in terms of the typical student rating form. Even within the domain of classroom instruction, students are in a position to judge only the delivery of material, not whether the material is accurate, up to date, or academically worthwhile. It can be seen, then, that student ratings are intended to measure only a limited range of teaching performance, namely the "delivery" aspect of classroom instruction, and thus, must always be supplemented by other measures which provide information about other aspects of university teaching.

A second limitation of student ratings is that they have only a limited degree of reliability and validity, thus leaving considerable margin of error in individual cases. For example, ratings tend to have low reliabilities across courses, so that a teacher who received a low rating in one type of course might receive a considerably higher rating if given the opportunity to teach a different type of course. Firm conclusions cannot be made about teaching ability from evidence gathered in one type of course. Ratings are also affected by bias factors, such as class size. We can try to control this bias by using separate norms for different class sizes, but there is always going to be some degree of bias in any rating scale. All forms of person perception are affected by extraneous factors, by the personality of the rater, by the needs of the situation, and student ratings are no exception. A further problem is that ratings show only a modest correlation with student achievement, so that a teacher could have low ratings from students but nevertheless be very successful in terms of student achievement. Conversely, a teacher might receive high ratings, but be unsuccessful in terms of student achievement. It is impossible to detect these types of teachers if we rely solely on student ratings.

A third problem with student ratings is that they show excessive "leniency bias". In other words, students are unwilling to assign low ratings. This is considered a problem or limitation in the sense that it decreases the sensitivity of student ratings. The ability of student ratings to separate good teachers from poor teachers is reduced by the tendency of students to rate everyone as a good teacher. A rating scale whose midpoint indicates "average" performance should theoretically yield results where 50% of ratees score above average and 50% score below average. This is the nature of an average. But student ratings do not work this way. Students are willing to say that 80-90% of their teachers are "above average". This leniency bias decreases the sensitivity or the range of variation in student ratings. The situation could be corrected by better instructions to students, explicitly advising them to rate the teacher in comparison to other university teachers, and perhaps by using percentile cues such as "top 5%" or "bottom 20%".

A further limitation of student ratings is that they do not apply well to atypical courses, for example individual study courses, or studio art courses, or "Keller plan" courses, where the teacher does little or no didactic lecturing, and in fact may never come into face-to-face contact with the students. The problem with this type of course is that the teacher's contribution occurs behind the scenes and thus is not easily observed by students.

Yet another limitation is that student ratings can only make relative differentiations among teachers rather than an absolute diagnosis of who is an outstanding or an incompetent teacher. If someone receives the lowest rating in the department of 40 people it does not necessarily mean that that person is inadequate or incompetent as a teacher. Among 40 superb teachers, someone must be least superb. Student ratings can provide only relative or rank-order information about teachers, not absolute categorizations.

A sixth and final problem is that student ratings can lead to the "twin sins" of overinterpretation and complacency. As student ratings gain institutional acceptance there is an increasing tendency to believe that these ratings have the same accuracy of measurement as a micrometer. This type of overinterpretation is seen, for example, when a discrimination of practical importance is made between an instructor with a 3.64 rating and a fellow instructor with a 3.62 rating. Individuals responsible for interpreting student ratings in personnel decisions must constantly remind themselves that these ratings contain a large error component. Closely related to overinterpretation is the tendency of students, faculty, and administrators to assume that once a student rating program has been implemented, all problems of evaluating teaching have been solved, and nothing more needs to be done. My own university is very guilty of this type of complacency. Ten years after the introduction of student ratings we still have not yet developed the supplementary measures of teaching that are needed to provide a full assessment of faculty performance.

Use of student ratings for improvement of teaching

Do student ratings lead to improved teaching? Many observers believe that teaching improvement is the only justifiable reason for using student ratings of teaching. From a researcher's point of view, the important question is whether or not feedback from students really does lead to improved teaching. Results of studies of this question are not optimistic. Most studies have utilized an experimental group of teachers and a control group of teachers, each of which is rated by students at two points in time, usually the middle of the academic term and then again at the end of the term. Feedback in the form of mid-term ratings is given to the experimental group but withheld from the control group. It is expected that the experimental group should receive better end-of-term ratings than the control group, as a function of receiving mid-term feedback from students. Rarely has this result been shown conclusively. In most studies, student feedback makes little difference in subsequent ratings (e.g., Centra, 1973b). Perhaps this is not surprising, because the student rating forms used in these studies have been of a "global-evaluative" nature which would be appropriate for promotion and tenure purposes. While these rating forms may be appropriate for promotion and tenure, they have not been constructed to help someone improve their teaching. Being told you are unsatisfactory in "clarity" or "overall effectiveness" does not tell you exactly what is wrong or exactly what to do to improve your teaching. A proper test of the effects of student feedback should use specific diagnostic feedback about particular teaching behaviours, as well as specific suggestions for improvement. Only then does the teacher know what the problem is and what should be done about it. Student ratings on the typical rating form are not informative enough to provide adequate feedback and consequently research results have been largely negative.

But giving teachers specific, diagnostic feedback, in and of itself, is probably not a sufficient condition for improved teaching. A second requirement for student feedback to be effective is that there must be motivation to improve (cf., McKeachie, 1976). As I suggested earlier, one advantage of incorporating teaching evaluations into the promotion and tenure system is that this practice gives faculty members a tangible incentive for improving their teaching. I have collected some data in my own department which bears on this issue. I located a sample of 30 faculty members who had taught in my department for five consecutive years between

1969 and 1979. I wanted to see whether their teacher ratings were improving from one year to the next, as a result of increased experience and regular feedback from students. At first, I computed the mean rating in each consecutive year for all 30 teachers. The results showed no improvement over time. Of course my sample included people both with and without tenure. One might assume greater motivation to improve teaching in the group without tenure, for obvious reasons. When the sample is divided into two subgroups, those with (N = 18) and those without (N = 12) tenure, the results are very interesting, as seen in Figure 3. Those without tenure showed steady improvement in their teacher ratings, whereas those with tenure showed no overall improvement from Year 1 to Year 5. These results suggest that feedback from student ratings leads to improvement in teaching only if the motivation to improve is sufficiently compelling.

The final ingredient needed for student feedback to be effective in improving teaching is that teachers must know what the correct alternatives are and be capable of implementing them. Many teachers fail to improve, despite diagnostic feedback and appropriate motivation, because they are either unaware of more effective teaching strategies, or unable to implement these strategies. This type of teacher not only needs to have feedback, but also some kind of training procedure, workshop, coaching, or other form of support for teaching improvement. In many universities, little or nothing is available in the way of intensive teaching improvement programs, other than the traditional "physician heal thyself" program. Even when universities institute formal programs

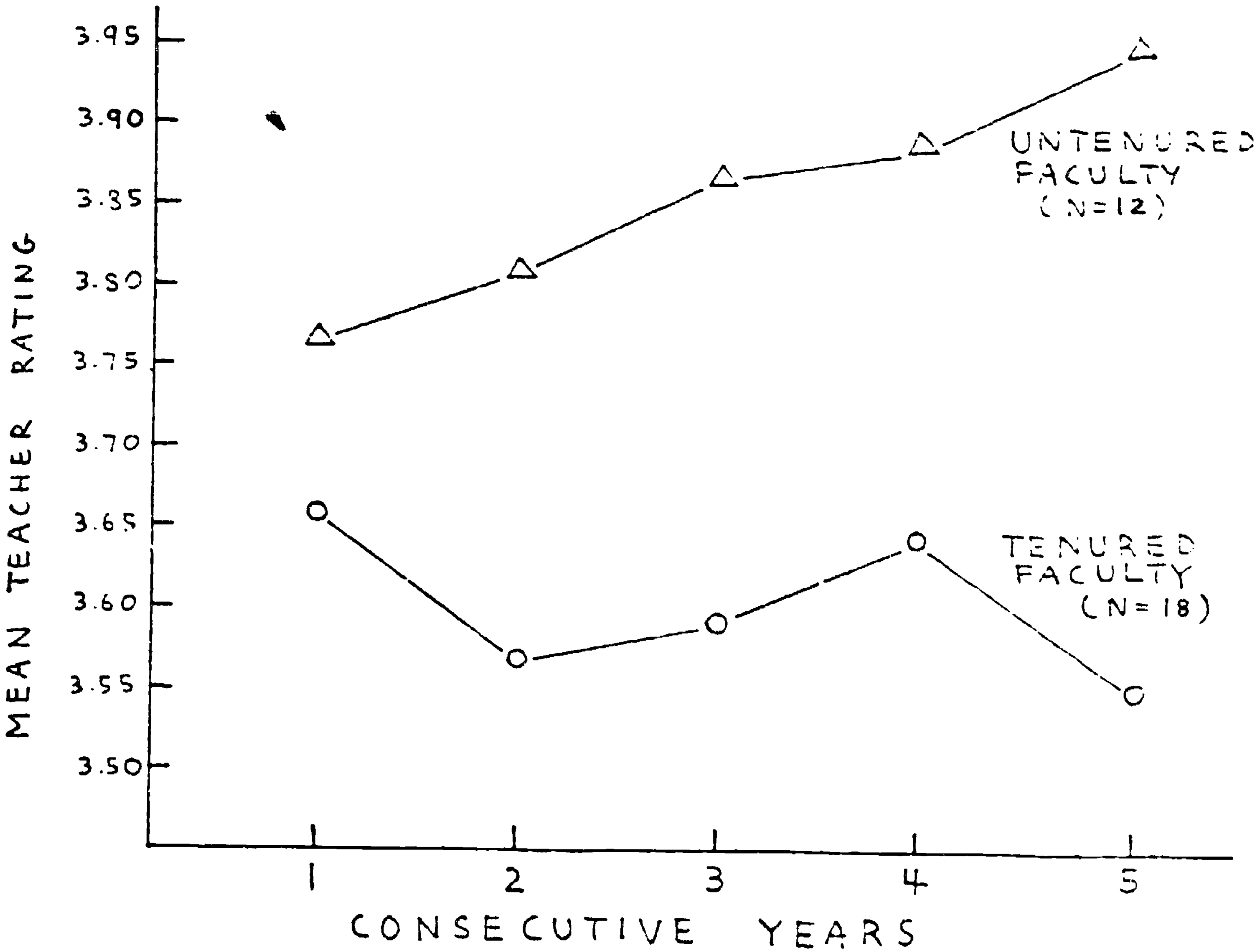


Figure 3. Mean teacher ratings for faculty members who did or did not have tenure over a period of five consecutive academic years.

or centres for teaching improvement, they often tend to be small-scale, poorly funded operations which have little or no chance of success.

In summary, I think it is true that student ratings can improve teaching if certain conditions are met. The ratings must be informative and diagnostic, there must be tangible incentives for improvement, and teachers need to have practice or coaching in implementing the correct alternatives. If these three conditions are met, student ratings can lead to improved teaching. To use global ratings alone and expect these ratings to lead to improved teaching is naive, however.

Alternatives to student ratings

In closing, I want to say something about alternatives to student ratings in the evaluation of teaching. The two main alternatives are (1) direct measures of student achievement and (2) colleague evaluations. I believe that these measures should be used as supplements to student ratings, but I am convinced that neither in isolation is as defensible a measure as student ratings.

Let me first consider direct measurement of student learning. To say that the best teacher is the one whose students learn the most, and the worst teacher is the one whose students learn the least, has intuitive appeal. It is easy to agree with a statement like this, but almost impossible to put into action. For purposes of illustration, assume that we wish to compare two teachers, one offering a fourth year honours seminar, and the other a first year lecture course, in terms of how much students learn. Since they teach different subject matters the only way we can compare these teachers is in terms of how successfully they achieve their own teaching objectives. Of course, it is unlikely that they would state their teaching objectives. University professors do not do this sort of thing, as we all know. Trying to get university professors to state their teaching objectives is like getting blood from a stone. But let us be wildly optimistic and assume that we have a statement of specific objectives from each teacher. We could then give students pretests and posttests to determine how much they have gained in the achievement of these objectives from the beginning to the end of the course. Let us suppose that students show a 50% gain in knowledge of subject matter in the first year course, and a 20% gain in knowledge in the fourth year course. Can we conclude that the first year teacher is more effective? Well, not really. It may be that a 20% gain score in a fourth year course is actually a better result than a 50% gain in a first year course. We have no idea what are the upper and lower bounds on these measures. The only way we can make a decision is by comparing each teacher to a set of norms for other teachers offering similar types of courses at the same or at other universities. Unfortunately, the expense involved in developing instruments and accumulating data for an operation of this sort would be beyond the budget of any university with which I am familiar.

Probably the most feasible situation in which to assess teachers in terms of student achievement is in a multiple-section course with a common final examination. But even here there are major problems. For one thing, it would be necessary to use random assignment of students to class sections to ensure that sections are equal in terms of student ability. Only then would it be fair to compare one teacher against another in terms of student performance on the common final exam. Rarely is there anything approaching random assignment of students to class sections in actual practice. A further problem with using final examination performance as a measure of teaching effectiveness is that this measure shows low reliability across years. In other words, if Professor X's students do well on a final examination in one year, they may or may not do well on a similar examination in the same course in another year. The correlation across years for this type of measure is only in the range of .30 to .40, not high at all, whereas student ratings of teaching have reliabilities of .70 to .80 in the same situation.

A second alternative to student ratings of teaching is colleague evaluation. There are three things colleagues might do in evaluating university teaching. First, they can watch teaching in the classroom and give their opinion about the quality of what they see. But students can provide adequate information on classroom teaching, and the number of colleague raters (and visits per rater) needed to achieve acceptable levels of reliability would be very expensive (Centra, 1975). And even if reliable colleague ratings were obtained, it seems likely that they would correlate closely with student ratings of classroom teaching. Thus it does not seem to be worth the effort to send colleagues into the classroom. The information obtained would be very costly and would probably be redundant with that obtained much more inexpensively from students.

The second role for colleagues is perhaps more promising. Colleagues can evaluate factors such as teacher knowledgeability and quality of course content, which students are not in a good position to observe. Colleagues can decide whether the course content is accurate, worthwhile, and up to date, and whether the course has proper academic standards. This can be judged from inspection of course materials, including textbooks, reading assignments, objectives, course outlines, and final examinations. Certainly this is useful input on teaching and is something that is not provided by students.

The third thing that colleagues can do is totally ignored in most teaching evaluation programs, but important nonetheless. I am referring to the assessment of non-classroom teaching activities such as course design, preparation of educational media or materials, policy making, and supervision of theses. Teaching activities occurring outside the classroom are events we should pay more attention to, and colleagues are in a better position than students to observe most of these activities. If we combine student evaluation of classroom teaching with colleague evaluation of course materials and non-classroom teaching, and with any defensible measures of student learning that are available, then we are at least getting closer to a full assessment of university teaching.

References

Bejar, I. I. A survey of selected administrative practices supporting student evaluation of instructional programs. Research in Higher Education, 1975, 3, 77-86.

Brown, D. L. Faculty ratings and student grades: A university-wide multiple regression analysis. Journal of Educational Psychology, 1976, 68, 573-578.

Centra, J. A. Reliability of student Instructional Report items. SIR Report No. 3. Princeton, New Jersey: Educational Testing Service, 1973 (a).

Centra, J. A. The effectiveness of student feedback in modifying college instruction. Journal of Educational Psychology, 1973, 65, 395-401 (b).

Centra, J. A. The relationship between student and alumni ratings of teaching. Educational and Psychological Measurement, 1974, 34, 321-325.

Centra, J. A. Colleagues as raters of classroom instruction. Journal of Higher Education, 1975, 46, 327-337.

Centra, J. A. Student ratings of instruction and their relationship to student learning. American Educational Research Journal, 1977, 14, 17-24 (b).

Centra, J. A., & Creech, F. R. The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness. SIR Report No. 4. Princeton, New Jersey: Educational Testing Service, 1976.

Crittenden, K. S., Norr, J. L., & LeBailly, R. K. Size of university classes and student evaluation of teaching. Journal of Higher Education, 1975, 46, 461-470.

Doyle, K. O., Jr. Student evaluation of instruction. Lexington Massachusetts: D. C. Heath, 1975.

Doyle, K. O., Jr., & Crichton, L. I. Student, peer, and self evaluation of college instructors. Journal of Educational Psychology, 1978, 70, 815-826.

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

Frey, P. W. A two-dimensional analysis of student ratings of instruction. Research in Higher Education, 1978, 9, 69-91.

Glasman, N. S., & Gmelch, W. H. Purposes of evaluation of university instructors: definitions, delineations and dimensions. Canadian Journal of Higher Education, 1976, 6, 37-55.

Goldstein, R. J., & Anderson, R. C. Attitudes of faculty toward teaching. Improving College and University Teaching, 1977, 25, 110-111.

Hogan, T. P. Similarity of student ratings across instructors, courses, and time. Research in Higher Education, 1973, 1, 149-154.

Kohlan, R. G. A comparison of faculty evaluations early and late in the course. Journal of Higher Education, 1973, 44, 587-594.

Leventhal, L., Abrami, P. C., & Perry, R. P. Do teacher rating forms reveal as much about students as about teachers? Journal of Educational Psycnology, 1976, 68, 441-445.

McKeachie, W. J. Psychology in America's bicentennial year. American Psychologist, 1976, 31, 819-833.

Murray, H. G. The validity of student ratings of faculty teaching ability. Paper presented at Canadian Psychological Association meetings, 1972.

Murray, H. G. How do good teachers teach? An observational study of the classroom teaching behaviours of Social Science professors receiving low, medium, and high teacher ratings. Ontario Universities Program for Instructional Development Newsletter, February, 1977.

Murray, H. G. Evaluating university teaching: a review of research. Toronto: Ontario Confederation of Faculty Associations, 1979.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. The Dr. Fox lecture: a paradigm of educational seduction. Journal of Medical Education, 1973, 48, 630-635.

Rich, H. E. Attitudes of college and university faculty toward the use of student evaluation. Educational Research Quarterly, 1976, 3, 17-28.

Rodin, N., & Rodin, B. Student evaluations of teachers. Science, 1972, 177, 1164-1166.

Seiler, L. H., Weybright, L. D., & Stang, D. J. How useful are published evaluations to students selection courses and instructors? Teaching of Psychology, 1977, 4, 174-177.

Sullivan, A. M., & Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 584-590.

Ware, J. E., Jr., & Williams, R. G. The Dr. Fox effect: a study of lecturer effectiveness and ratings of instruction. Journal of Medical Education, 1975, 50, 149-156.