December 1996

On the Use of Numerically Scored Student Evaluations of Faculty

William Rundell

Introduction.

First, some historical background. Up until 1994 the Department of Mathematics at Texas A&M University relied on a student evaluation of instructor form that was heavily weighted towards verbal responses. Although there were questions such as "On a scale of one to five, rate your instructor as to ...", much more attention was paid to the verbal comments when faculty teaching was being evaluated, in particularly for promotion or tenure decisions. In any case, there was no tabulation of the results and certainly no attempt made to reduce the output from the questionnaires to any quantitative value.

For a two year trial period a decision was made to utilize questionnaires that were electronically scanned and the output easily condensed to a few numbers that were (hopefully) indicative of the teaching performance of the instructor, albeit from the student's point of view.

From an administrative viewpoint, there is an obvious lure to measuring "teaching effectiveness" by reduction to a few, or as actually seems to happen, a single number. Most often this number is the average of all the individual components. This is especially true when large numbers are involved. Reading 10,000 student questionnaires each semester for 115 faculty and putting this into any kind of context is virtually impossible. Even for the relatively few cases where in-depth information must be acquired, it is often difficult for evaluators to use solely written comments and yet put them in a comparative perspective. Thus, "Jones had a 3.94 mean on her student evaluations, and since this is 0.2 above the average for the Department, we conclude she is an above average instructor as judged by these questionnaires" is a statement that appears to be increasingly common. Since few departments have managed to similarly quantify other possible measures of evaluation, the single number stands out and eventually becomes *the* measure of teaching effectiveness.

The purpose of this article is to examine some of the issues surrounding this method of analysing the data; to come to some understanding of the information obtained and to determine whether the distillation down to a few numbers is a valid tool for the effective evaluation of teaching. In short, we want to see if what seems to be an increasingly common system of evaluation contains sufficient information of reliable nature as to be useful as a means of determining raises, promotions and tenure.

The Evaluation Form Used.

The main part of the evaluation form consisted of ten questions with choices ranging from 1 (low) to 5 (high). There was also a space for comments. This format is much more in line with that used by other departments in the University, and indeed the first five questions on the form were ones suggested jointly by the Student and Faculty Senates. The other five questions were added by a department committee and were meant to complement the other five. These questions are shown in Figure 1.

Figure 1. The Questions used

1. I would take another course from this professor.
2. The exams/projects were presented and graded fairly.
3. The amount of work and/or reading was reasonable for the credit hours received in the course.
4. I believe this instructor was an effective teacher.
5. Help was readily available for questions and/or homework outside of class.
6. The instructor seemed to be well-prepared for class.
7. The instructor had control of the class' direction.
8. The instructor genuinely tried to help the students learn the material and showed concern.
9. The instructor covered all the material and paced it evenly.
10. Compared with all the instructors I have had in college, this instructor was one of the best.

There was some awareness of possible overlap, for example, questions one and ten, but that was not considered critical. Typically, these evaluations were filled out in the last week of class, but prior to the final examination The questionnaires were processed by a central measurement and testing service. Raw output consisted of n words, each of length ten and consisting of an alphabet of the characters 1, 2, 3, 4 and 5. Here n denotes the total number of responses. Published output was the averages for each section of the responses to each of the ten questions, plus the average of these ten numbers, the "overall mean." Note that some questions pertained to the instructor while others to the course content. These are frequently linked, but in the case of a department that carries a large service load with multiple sections of a given course the instructor tends to have very little input into the course syllabus. Also available was the the answer to an eleventh question, the student's expected grade (this was not a standard feature of these forms but was added by the department), the total number enrolled in the section, N, the number of students who "Q-dropped", and, from a separate source, the grade distribution for that section. Of course, $n \leq N$ and the response rate n/N varied markedly. The mean response rate over all sections was 67.5% with a standard deviation of 16.3%.

Background and types of Course.

At Texas A&M we teach very little precalculus mathematics or material normally considered to be part of the high school curriculum. These courses account for about 6% of our enrollment. There is a core curriculum requirement, typically satisfied by a course in finite mathematics (M166) and one in "single semester calculus" (M131) and usually taken by those students whose major does not require a specific mathematics sequence. There is also a version of these, at much the same level of sophistication and content, but tailored towards the needs of those students enrolled in the College of Business (M141, M142). Average total enrollment in these four courses is 5,200 students in the fall semester and 4,500 in the spring. In these courses there is no implied sequencing and students frequently take them in any order. Although they have a freshmen designation, and from a mathematical standpoint that is correct, many students postpone these classes until later in their undergraduate program. The average class size is 100. For the purposes of this study we will designate the above classes as constituting Group A.

There is a three semester calculus sequence for Mathematics and Science majors and a parallel, virtually identical in content, one for engineering students. Average enrollment in these sequences is 3,100 students in the fall and 2,600 in the spring. Class size varies from about 30 in the mathematics/science sequence to a little little under a hundred students in the engineering sequence although the latter has recitation sections of average class size 30. We will designate these courses as Group B.

Almost all of the students in Group B take a fourth semester class in differential equations and most of them take at least one further course in mathematics at the junior level. The differential equations course has an average enrollment of 700 per semester and the other junior level courses have another 400. Class size is between 40 and 50 students. We will refer to these courses as being in Group C.

We also teach a variety of courses at the junior and senior level primarily designed for mathematics majors; these will be denoted by group D.

The multi-section courses we teach have a varying degrees and methods of coordination. In an experiment started three years ago the Department used common exams in the first two semesters of engineering calculus. In fact the actual questions posed are not known to the instructors of the class prior to the examination.

Of course, the Department teaches a wide variety of classes and the results from some of these will appear in the analysis below. However, the sheer numbers afforded by the enrollments in Groups A, B and C allow for an analysis of the data with minimum subjectivity. The data for this paper was accumulated over the four semesters the form was in use.

The Information Content of the Data

The first thing we might want to inquire after hearing of Professor Jones' "above average" student questionnaires is the class taught, for as the table below shows not all levels of student tend to work on the same scale.

Table 1. Mean Evaluation Norm for Professorial Rank Faculty

Course Type	Mean Evaluation
Group A	3.35
Group B	3.81
Group C	3.91
Group D	4.27
Graduate	4.47

This is certainly not unexpected since students who take a class as part of a mandated general education requirement might not have the same notions as a senior mathematics major or a graduate student. There are some flaws in interpreting the vast differences in mean scores in this table as being totally due to differences in the students. Both the undergraduate and graduate program committees succesfully lobby for certain instructors to teach the courses for our majors so that the professors involved do not constitute a random sample.

An intelligent use of data would surely take this into account. Would it also need to allow for the following?

1. "On semester" classes tend to have better students and higher grades than those in the "off semester." For example, science and engineering students are scheduled to take first semester calculus during their first fall semester, those who are considered not sufficiently prepared delay it to the spring and join those who are retaking the course from the fall.
2. The time of day the class was taken. For example, 8am and late afternoon classes fill slower and those during the mid-morning periods fill the quickest. Does that suggest a different type of student in terms of how he or she evaluates is forced eventually to take a less popular time period? The average grade point ratio given out in 8am sections is consistently lower than other time periods, especially when compared to mid-morning sections.
3. The percentage of students responding to the survey. As noted earlier, this varies widely between sections of the course, and as we will see later, there is evidence that the demographics of those responding to the survey is not representative of the sample. How should we compensate for this effect, if at all?
4. The percentage of Q-drops varies significantly between sections and this is true at all levels of course. For the Department the mean Q-drop rate is 8.34% with a standard deviation of 3.84%.

   One has to interpret a Q-drop as registering some dissatisfaction although it is difficult to assess the impact of this on the evaluation since these students did not participate in the process. Did instructor A who received a mean rating of 4.3 but had a 20% of his students Q-drop achieve a better "customer satisfaction" norm then instructor B who had no Q-drops but a mean rating of 3.9? Professor A's mean rating puts him into the well above category, while B's places her at slightly below the department average. If you make the assumption that those students who dropped would have given out a 2.0 then this would adjust the mean rating of instructor A to 3.84. Is this value of 2.0 the correct one to use?

   Would you expect a correlation between the Q-drop rates and the mean evaluation? On the one hand, one might expect that sections with a high Q-drop rate would have higher evaluation on average since some of those students who were dissatisfied were not part of the survey. One could alternatively argue that popular instructors would automatically have a lower Q-drop rate. In fact there is some evidence that the latter factor is the slightly more dominant one. For instructors whose evaluation was in the bottom one third of the range there was no evidence of a trend in either direction, at least as far as the gross data was concerned, the correlation index being -0.02. For the instructors who were evaluated as being in the middle third the correlation was -0.15 and for those in the upper third the correlation was -0.22.

A fundamental question is whether the students are answering the questions asked, or are the responses to individual questions coloured by their overall impression of the instructor, or the course, and if this the case, how strongly.

In order to quantify this we used two measures.

> Each individual questionnaire provides a 10 digit word, $x = x_1, x_2, \ \ldots \ , x_{10}$. The total lexigraphical distance between two words x and y is $d(x, y) = \sum_1^{10} |x_i - y_i|$. Perfect correlation for a given student's responses would have all characters in the word constant. Since each x_k lies between 1 and 5, completely uncorrelated responses based on random input would have a average distance of 2 per character or 20 per word.

> For each section the average scores on each question are tabulated, giving numbers between 1.0 and 5.0. The set of numbers corresponding to questions x and y are compared by the usual correlation formula for the correlation index for two pairs of quantities (x_k, y_k),

$$r = \frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_k (x_k - \bar{x})^2}\sqrt{\sum_k (y_k - \bar{y})^2}}$$

> Of course these measures are related but they do offer different ways to look at the situation. There are other norms we could have used; in particular, we could have worked with a correlation index based on the ℓ^1 rather than the ℓ^2 norm. This would minimize the effect of outliers.

Looking at the responses, either as the raw data from each evaluation within a given section or as the averages for each section, one is immediately struck by the high degree of correlation to the answers to each question.

For the questionnaires in group A the distance between words and their mean values was 0.57, meaning that it would require on average less than 6 changes by a single digit to make the response a constant one. In fact, 12.5% of all questionnaires were identically constant and more than a third differed by only three digits from a constant. This degree of uniformity actually increased with the sophistication of the course. For group C 15% of the students responded with a constant response and the mean distance from a constant response was 0.49.

The correlation between the various questions for the averaged responses from each section is shown in the tables below.

Table 2. Correlation Matrix for Group A*

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
Q1	1.00	0.85	0.81	0.97	0.85	0.74	0.73	0.91	0.83	0.98
Q2		1.00	0.85	0.80	0.76	0.64	0.68	0.85	0.80	0.85
Q3			1.00	0.75	0.75	0.61	0.58	0.80	0.86	0.78
Q4				1.00	0.85	0.83	0.79	0.90	0.83	0.98
Q5					1.00	0.79	0.68	0.92	0.79	0.84
Q6						1.00	0.80	0.79	0.76	0.78
Q7							1.00	0.67	0.68	0.76
Q8								1.00	0.83	0.91
Q9									1.00	0.83
Q10										1.00

* Number of sections correlated = 216 (12,649 students).
Average constancy of response = 0.578

Table 3. Correlation Matrix for Group B*

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
Q1	1.0	0.68	0.74	0.97	0.81	0.83	0.80	0.92	0.83	0.98
Q2		1.0	0.71	0.67	0.66	0.61	0.51	0.69	0.51	0.68
Q3			1.0	0.68	0.63	0.55	0.46	0.71	0.66	0.73
Q4				1.0	0.82	0.89	0.83	0.90	0.83	0.97
Q5					1.0	0.76	0.63	0.84	0.70	0.83
Q6						1.0	0.87	0.76	0.75	0.83
Q7							1.0	0.70	0.69	0.78
Q8								1.0	0.78	0.93
Q9									1.0	0.83
Q10										1.00

* Number of sections correlated = 176 (7,301 students).
Average constancy of response was 0.551.

Table 4. Correlation Matrix for Group C*

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
Q1	1.00	0.77	0.68	0.95	0.64	0.67	0.67	0.84	0.64	0.97
Q2		1.00	0.62	0.72	0.55	0.54	0.57	0.67	0.62	0.77
Q3			1.00	0.61	0.42	0.37	0.50	0.56	0.52	0.63
Q4				1.00	0.65	0.80	0.79	0.87	0.64	0.96
Q5					1.00	0.66	0.57	0.74	0.53	0.69
Q6						1.00	0.78	0.68	0.63	0.75
Q7							1.00	0.73	0.57	0.75
Q8								1.00	0.65	0.86
Q9									1.00	0.68
Q10										1.00

* Number of sections responding was 95 (2,878 students).
Average constancy of response was 0.491.

Table 5. Correlation Matrix for Group D*

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
Q1	1.00	0.75	0.69	0.95	0.53	0.68	0.61	0.74	0.69	0.94
Q2		1.00	0.73	0.66	0.52	0.41	0.37	0.67	0.56	0.69
Q3			1.00	0.64	0.40	0.49	0.53	0.51	0.55	0.61
Q4				1.00	0.47	0.75	0.65	0.66	0.68	0.95
Q5					1.00	0.37	0.31	0.77	0.49	0.58
Q6						1.00	0.79	0.44	0.67	0.79
Q7							1.00	0.26	0.58	0.64
Q8								1.00	0.60	0.73
Q9									1.00	0.77
Q10										1.00

* Correlation of 46 junior/senior mathematics majors sections (831 total responses).
Average constancy of response was 0.448.

Are we really to believe that the correlation to questions 1 and 2, or 1 and 3, one designed to gain information about the instructor in general and the other specifically about the course content, should be so high? Even in those sections of freshmen engineering calculus with common exams where the instructor had no input into the course syllabus or either the making or grading of the major exams the correlation between these two questions was considerable (0.71).

While there is very clearly some signal remaining from the responses to individual questions, it is just as obvious that it is blurred by a background that in some sense measures the student's attitude towards the instructor, what one might call the "customer satisfaction index." It is also clear that the signal to background ratio is quite low and this shows that caution must be used in interpreting the responses as legitimate answers to individual questions.

The above correlations might give credence to the practice of replacement of the individual response averages by their mean. On the basis of the above information this is certainly justified from a statistical standpoint. It does beg the issue of why we should bother to have ten or more questions when two or three might very well carry virtually the same information. It also motivates the further study of the student evaluation data in order to see on what other factors it might depend.

Evaluation and Grades.

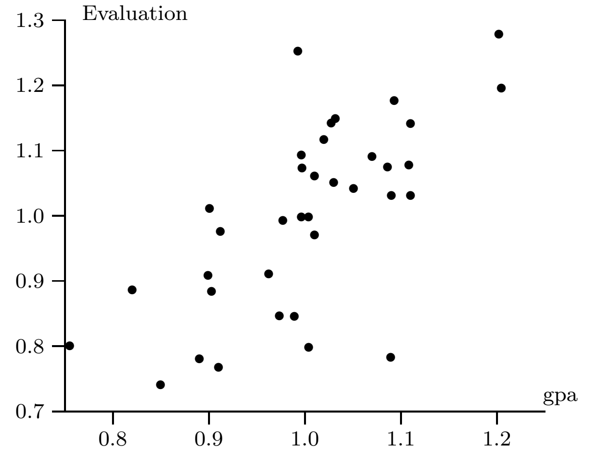
Folklore in the Department, and indeed amongst mathematics faculty nationwide, has long held that there is a direct correlation between student evaluations and grades, despite an extensive claim in the Education literature to the contrary. The data we have accumulated over the last three semesters allows us to make some tests of these contrary hypotheses.

The figures to be presented below plot the gradepoint ratio given out in each section against the mean evaluation score for that section. Care must be taken in selecting the courses and choosing the scales. We avoided courses that had strong coordination between sections, such as a common exam format since these, at least in theory, should have no instructor-dependent variation in grades (although the laboratory and final exam grade were at the instructor's discretion). On the other hand we wished to choose courses with a large number of sections in order to get statistically meaningful results. We cannot compare sections of different classes on the same graph for they are very likely to have different gradepoints and possibly different averages for the evaluations, as table 1 shows.

We also should be aware of the issue of comparing sections of a course on the "on semester" with those on the "off-semester" since the grade scales are going to be different. We can allow for this last possibility by dividing individual section gpa's by the mean gpa for that semester. Thus a relative gpa of 0.95 meant that the gpa for that section was only 95% of the mean gpa for all sections of that course in that semester. The same sort of scaling can be done for the evaluations. The figures below show the results for three courses that fit the above paradigm. In these figures we have used the following notation: For a vector of values $\{x_i\}_1^N$ we denote by $|x-\bar{x}|_1$ the quantity $|x-\bar{x}|_1 = \sum_1^N |x_i - \bar{x}|$. Given two sets of data values $\{x_i\}_1^N$, $\{y_i\}_1^N$ we will make the hypothesis that they obey a linear relationship of the form $y = ax + b$. The quantities a and b are computed by a least squares fit to the data (x_i, y_i). In this situation σ_1 is the sum of the absolute values of the distances of each point (x_i, y_i) from this line.

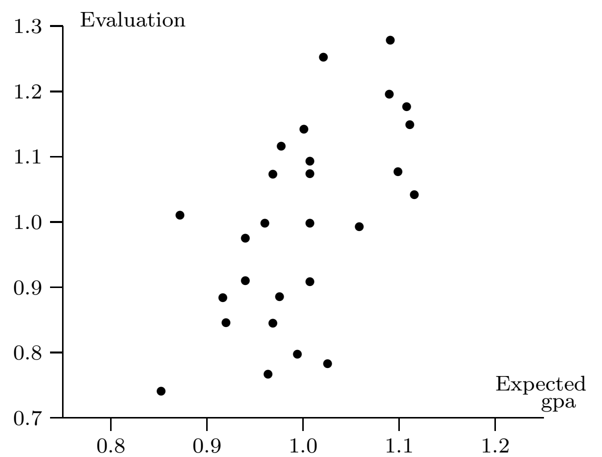
Figures 2, 4 and 5 below show plots of the actual grade point ratio given out against the total evaluation for that class for each of one course selected from Groups A, B and C. Figure 3 shows an example of a student's expected grade plotted against total evaluation. The influence of Q-drops has not been taken into account in these figures. Some discussion of the influence of this factor will be presented later.

Fig 2. Math 251: actual gpa and evaluations



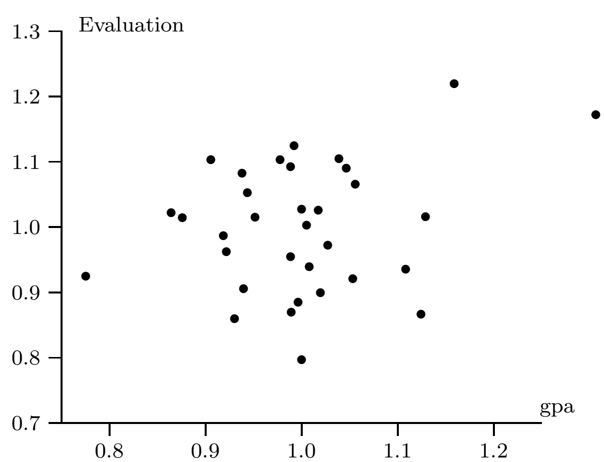
$|x-\bar{x}|_1 = 0.073$ $\sigma_1 = 0.077$
$|y-\bar{y}|_1 = 0.117$ r $= 0.675$

Fig 3. Math 251: expected gpa and evaluations



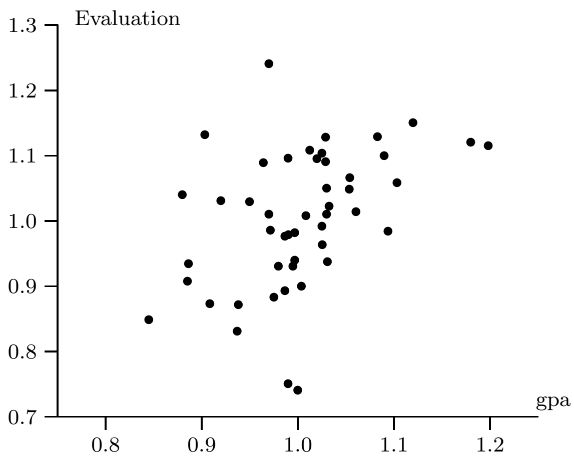
$|x-\bar{x}|_1 = 0.056$ $\sigma_1 = 0.127$
$|y-\bar{y}|_1 = 0.127$ r $= 0.602$

Fig 4. Math 142: actual gpa and evaluations



$|x-\bar{x}|_1 = 0.067$ $\sigma_1 = 0.078$
$|y-\bar{y}|_1 = 0.081$ r $= 0.270$

Fig 5. Math 308: actual gpa and evaluations



$|x - \bar{x}|_1 = 0.053$ $\sigma_1 = 0.070$
$|y - \bar{y}|_1 = 0.084$ r = 0.439

These correlations are too high to accept the hypothesis that grades and evaluations are unrelated.

The figures presented here indicate a possible, albeit crude, means of taking into account the influence of grades. Given a certain degree of correlation has been detected, one can make the assumption that points close to the line $y = ax + b$ of best fit are within the "expected range." Points (x_i, y_i) that are outliers and above this line represent those instructors who have significantly higher evaluations than the values of the grades given out in this course. The individuals in this category correlate very strongly to those previously considered by the Department to be "good" instructors. Indeed, this is where many of the winners of teaching awards placed. In a similar manner it was noticed that instructors whose scores placed them significantly below the line had previous histories of poor student relations.

To further pursue the relationship between evaluation responses and grades, we looked at each group of courses and computed the mean score for each of question 1 through 10 in the case when the expected grade as provided by the student on the form was each of A – F. The results are shown in Table 6 below for groups A, B and C.

Table 6. Evaluation against expected grade

	Expected Grade	Average Eval Question 1	% claiming this grade
	A	3.855	24.9
	B	3.491	37.0
Group A	C	3.063	26.9
	D	2.592	6.7
	F	2.415	0.5
	A	4.094	24.5
	B	3.730	39.5
Group B	C	3.299	28.0
	D	3.127	4.7
	F	2.6.97	0.7
	A	4.171	29.8
	B	3.695	40.3
Group C	C	3.217	22.4
	D	2.704	2.6
	F	2.375	0.4

Similar values were obtained for every sufficiently large sample we analysed.

Using the data in this table the correlation between the average evaluations and the expected grades is 0.99 although this is a misleadingly high figure since we have averaged out much of the detailed information of the raw data, Another way to look at this information is this. For the evaluations in group C, students who expected an A were four times more likely to award the instructor a score of 5 in question 1 than those students whose expected grade was a C. Similarily, the C students were four times more likely to respond with a score of 1 than the students who expected an A.

It should be noted here that the number of D and F grades claimed by the students as their expected outcome are considerably lower than actually given out for the class as a whole. The other grade values, while more in line with actuality, are over-inflated. There are several possibilities that might explain this. The first is that the students simply overestimate their actual grade. They either respond here with an unrealistic expectation or the final exam does consistently lower grades. Second, the expected grades claimed are approximately correct, and the missing students in the evaluation process are mostly making grades of D and F. If we note that that on average there is only a two-thirds response rate, then this second hypothesis is not entirely inconsistent with the data. The likely situation is a combination of these possibilities with some variation depending on the course. For example, courses for elementary education majors had a higher difference of actual from expected grade point ratios despite having one of the highest response rates of any group.

Evaluation and Future performance.

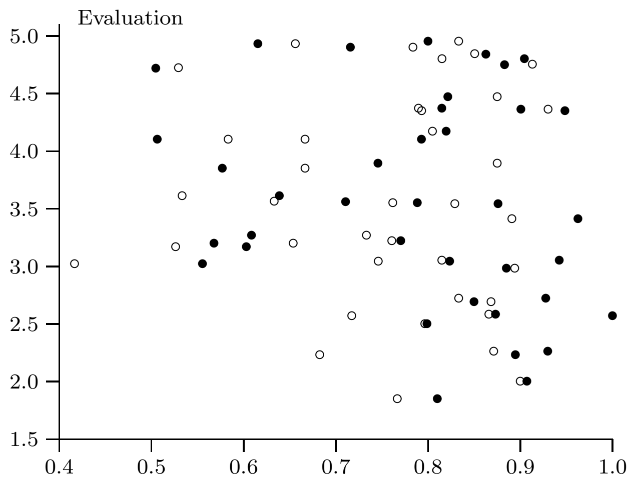
We made an attempt to use some of our longer sequences of courses to determine whether students who took certain instructors in the beginning courses of the sequence did measurably better than others in later courses. The aim was then to test the correlation between "successful professors" (by this measure) and professors who had "good" student evaluations. The group of courses selected for the experiment were the three semesters of calculus and differential equations. No other sequence we teach is as long and contains as many students as this one.

One always hopes for internal consistency and some positive correlation between the classes of "good instructors" by two different measures is to be expected. While this sample is more limited than we would like, we did not use any information consisting of less than 25 students and this is the minimum sample size for each of the dots in the graphs below. The mean sample size was approximately 50.

There are several ways to define success in successor courses, and in the data presented below we used two such measures. One was simply making a grade of C or better in the more advanced class given that a passing grade was achieved in the lower one. The other one looked at the grade point ratio of each of the cohort groups and defined success as the ratio of the gpa in the advanced class divided by the gpa of the group in the lower course. In the first case, for every student in common who made a C or above one point was given, for those who made a D or F no points were assigned. Thus the average score by this method was always less than or equal to one. For the second case, the ratio could certainly exceed one, but as a testimony to the high standards required of students in this sequence, this was rarely true.

The case of the first semester (M151) as a feeder for the the second semester (M152) is shown in Figure 6 below. The points designated by a ○ correspond to those where success is measured by making greater than a grade of C, and the points ● denote the ratio of successive gpa's

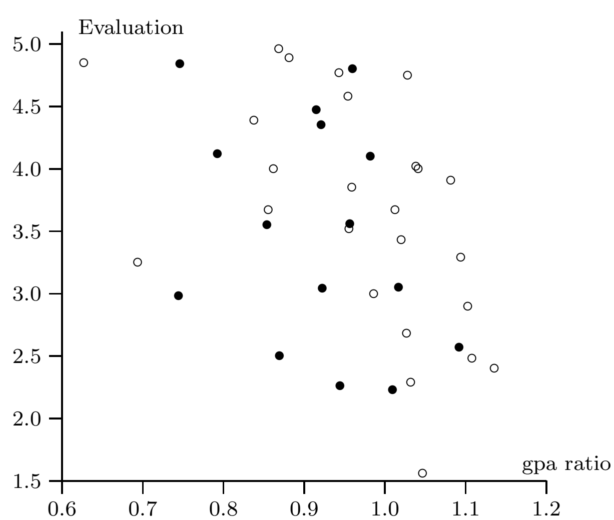
Fig 6. Math 151→152 success rate *vs* Average Evaluation



● gpa ratio (r = -0.240)
○ Proportion making ≥ C (r = -0.030)
sample size = 1887

Figure 7 shows the situation with the advanced class being third semester calculus and the feeders being either first or second semester calculus. The success criterion is the gpa ratios and this is compared to the response to question 1. If we had to use the greater than grade C criterion the corresponding correlations are $r = -0.367$ for M151 and $r = -0.160$ for M152.

Fig 7. Math 151, 152→251 success rate
***vs* response to Question 1**



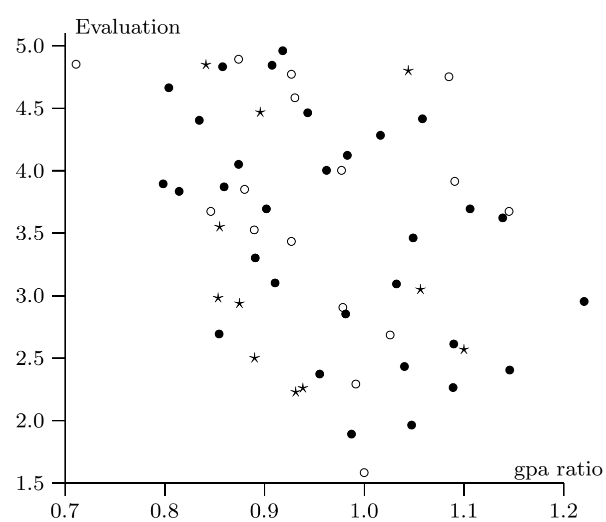
● from M151 (629 students) r = -0.343
○ from M152 (1083 students) r = -0.509

The most charitable way to describe this information is that it shows by some measures the two quantities are almost unrelated. A more cynical voice would consider the negative correlation factor to be significant. In either case it is disturbing to note the number of professors who received very high student evaluations yet performed poorly on the carry-on success rate test. Many of the faculty who gained high scores on the carry-on success rate were instructors with a reputation for excellence, but who are considered "hard" by the students. Traditionists will consider this data to be a triumph for their viewpoint. The considerable range in values of the "success rate" parameter is certainly striking, and at least to the author, unexpected.

There are several criticisms that can be made in the computation of this success rate. Amongst these are the fact that we neglected Q-drops in computing the gpa ratios. An instructor who had a large number of Q-drops in the beginning of the two classes passed on a reduced, and presumably stronger, set of students than the instructor who retained a higher proportion of the intake. Also, a Q-drop in the upper level course might have some significance.

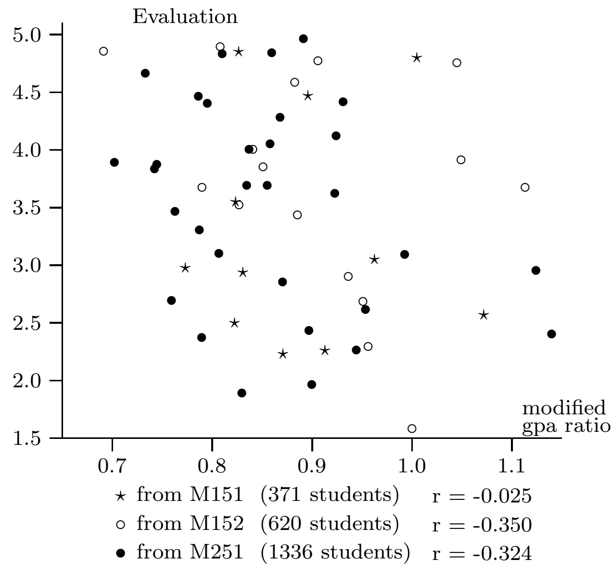
To test for these effects we used the differential equations class M308 as the final course and with each of the first, second and third semester calculus as the initial class but this time made two modifications to the grade-point ratio to reflect the Q-drops. In the final course (M308) we assigned a grade value of 1.0 to a grade of Q, while in the initial course we multiplied the grade point ratio of the students passed on by the retention rate (those completing the course divided by those officially registered). The possible effect of Q-drops on the evaluation scores themselves was not taken into consideration. The results are shown in Figure 8 which uses the actual grade point averages for the initial course and in Figure 9 where the gpa has been modified by the described process.

Fig 8. Math 151, 152, 251→308 success rate
***vs* Average Evaluation**



⋆ from M151 (351 students) r = -0.127
○ from M152 (589 students) r = -0.324
● from M251 (1242 students) r = -0.467

Fig 9. Math 151, 152, 251→308 modified rate
***vs* Average Evaluation**



The average Q-drop rates for the three feeder courses is not the same; 3.9% for M151, 4.3% for M152 and 9.0% for M251. As one can see, this modification for Q-drops makes some change in the in correlation index, but the overall picture is very similar. This is not to say that Q-drops should also play a small part in other correlations, and in particular, not for the case of an individual instructor. For example, when the data used to construct figure 6 was run to include Q-drops as a non-success factor very little difference was found for most instructors. This is consistent with the low rate of Q-drop in the second course (M152). However, a few instructors had their success rate drop by 50%, an indication of a possible problem.

An objection to using student performance in downstream courses as a mechanism for evaluating teaching effectiveness is that in longer chain sequences the instructor of the first course may be expected to have a lesser effect on the performance of his or her students on the final class in the sequence due to the influence of the intermediary instructors. More importantly, this may be difficult to quantify with the available amount of information. There is some evidence of such a difference from the data contained in figures 8 and 9.

Another objection to this method of evaluation is the possibility that only serious students willingly take courses from the "hard" instructors and so this faculty group start out with an unrepresentative sample. If this is in fact the case then it would indeed be difficult to compensate for the skewed sample. However, even with full information students are rarely able to plan their schedules to take every section of choice, and it is unlikely that this effect is any greater than that attributable to the time the class meets. However, the assumption that all sections of a course being compared have very similar students profiles as regards ability and motivation is critical to the analysis of this section. One facet of this particular issue will be studied in the next chapter. The results indicate that the assumption is in fact open to serious question.

The limited nature of this study must be understood. Only four semesters of information was available and only a relatively small number (40%) of the faculty participated in the this calculus sequence Yet there was considerable consistency in many respects. Instructors who appeared in more than one graph tended to have very similar carry-on success rates. However, it must be noted that this rate, being a single number, cannot be expected to capture many of the nuances of a complex process and most of the modifications suggested as being relevant for the average student evaluation index are applicable here also.

If this carry-on success rate is a valid measure then there must be legitimate concern with the use of student evaluations as a primary source of teaching effectiveness.

Do all sections have similar student profiles?

This assumption is implicit in any comparisons that might be made between different sections of a given course. It is particularly important to be able to make such an assumption in the study of success rate in downstream courses, or if it is invalid, to find a means of compensating for the effect.

Attempting to test motivation levels of students is beyond any data set that the Department possesses. However, we can test whether the students in two different sections have obtained a similar level of academic performance. For mid or upper level courses we can use the current grade point ratio of the students. This will not be useful for freshmen, but we can use an index such as SAT score or high school rank. Note that this is not to indicate a believe in the strong predictive powers of standardized tests, but merely an acknowledgement that such measures are frequently used as predictors, and is data that we do have available.

To test the hypothesis that all sections of a given course tend to have similar student profiles we looked at the distribution of SAT scores in the first semester of engineering calculus (Math151) and the distribution of grade point ratios of students in the sophomore/junior level course Math308. This was done for all sections of these courses over a 6 year period, 1990-1996. In each semester the total number of students in each course were considered to be distributed into N groups according to their scores (SAT or GPA) and the qualifying scores for each of these groups were chosen so as to make approximately equal numbers in each group. We used a value of $N = 10$. In the case of Math151 the average section size was 100 students so that there is an expectation of 10 students in each level or bin. For Math308 the expected number in each bin is 6, but this is still more than the accepted minimum for the χ^2 test that was used. Sections with lower enrollments that did not meet a minimum expected level of 5 students in each bin were deleted from the study.

Using the χ^2 statistic the probability that the distribution into each of these N bins from any given section is fairly drawn from the course sample can be computed. This will be a number between 0 and 1 with low values indicating poor correlation. The table below shows the range of these values over all sections for the six year period.

Table 7. Distribution χ^2 probabilities obtained from comparing individual sections of a course with scores from all students in that course in a given semester.

χ^2 range	M151*	M308†
0.0 – 0.1	7.1%	4.0%
0.1 – 0.2	7.1%	4.0%
0.2 – 0.3	7.1%	3.0%
0.3 – 0.4	8.9%	4.0%
0.4 – 0.5	8.0%	8.0%
0.5 – 0.6	7.1%	11.0%
0.6 – 0.7	13.4%	12.0%
0.7 – 0.8	14.3%	13.0%
0.8 – 0.9	11.6%	21.0%
0.9 – 1.0	15.2%	22.0%

* Using SAT score as the comparison (112 sections).

† Using previous gpa as the comparison (101 sections).

This certainly does not support the hypothesis that test scores of students differ little between sections. A similar set of differences was found for others courses analysed.

Conclusions

We entered into the process of standarised evaluations with an open mind and were hoping, as many have in the past, for a silver bullet that would allow us to deal with the problem of evaluating teaching in an objective manner. If this could also be combined with a reduction in the workload of such a task then this was an added bonus.

There is much information that can be gained from the numerically-based responses and there is clearly a signal hidden in a background of more single-valued information. How to filter this background is much less clear. How to modify the responses in light of other information about the course is even less clear. A mathematical model can be constructed that tries to make allowance for these factors and the available data used to get a best fit to needed parameter values provided the sample size is sufficiently large. The complexity of this will be orders of magnitude greater than anything that is currently being attempted locally and there will still be significant subjectivity in the interpretation of the results.

However the analysis we have performed on the data suggests that the distillation of evaluations to a single number without taking into account the many other factors can be seriously misleading. The correlation between positive student evaluations and grades awarded is sufficiently strong to indicate that a procedure based on numerical scores such as we have described is surely going to lead to grade inflation in the long term.

While the idea of tracking student's progress through a sequence of courses is an attractive means of evaluating faculty performance, only a relatively small number of our enrollments in a given semester is in a chain of courses sufficiently structured for data to be collected. For various reasons some faculty do not teach these classes at all. The negative correlation that our study seems to indicate between the two measures of "carry-on success" and, what we have discovered in this article is best described as "short term customer satisfaction," is very disturbing. If this is indeed the situation, then the use of student evaluations as a primary measure of teaching effectiveness, simply because it is easily normable, is a very questionable practice.