Trust, but Verify: Toward Fairer Student Evaluations

by John Daley

In seven years of college teaching, I have encountered no less than six student-faculty rating instruments. Each has strengths and flaws — as one might expect — and, if objectively compared, there would be no outright winner or loser. But one thing is certain: My favorites are the ones that administrators have rejected the fastest.

Can student teaching evaluation instruments that satisfy students and administrators ever provide faculty with useful insights on their classroom performance?

Perhaps my naiveté is getting the better of me, but I believe the answer is still yes. I believe enough in the student evaluation's diagnostic potential to risk the impertinent and the downright nasty in order to glean a few useful insights. Besides, I have found that a good 90 per cent of the students who fill out evaluation forms do so honestly.

However, the risk of the other 10 per cent skewing the evaluation's results — because they were earning lower test scores than they would have liked is significant.

On the evaluation instrument currently used by our department, three students out of a section of 30 can lower that section's raw average 0.8 points on a scale of 8 by giving straight zeros.

Unfair evaluations can cause disparate scores for two sections of the same course taught by the same professor. In one of my own recent experiences, the disparity was over 50 percentile points in breadth. If most of the adverse evaluations in this instance had been based on objective observation or an honest reporting of subjective impressions, the percentile scores would likely have varied little from one section to the next.

How can we teachers diagnose our strengths and weaknesses using such instruments as these? If you cut through all the administrative newspeak about "responsiveness," the answer, simply put, is that we cannot.

John Daley *is an assistant professor at Pittsburg State University in southeastern Kansas, where he specializes in military and nineteenth century American history. One of three recipients of PSU's Outstanding Faculty Award for 1998-99, he is an active member of the local Kansas NEA chapter.*

Evaluators should seek out objective evidence that can shed light on the students' subjective impressions.

When we ask our administrators this question, however, we're told that "student empowerment will not hurt the 'good' teachers."

I have listened to this oft-repeated party line the same way I listen to a broken record — at first with mild annoyance and, then, with consummate frustration.

Considered against this increasingly adversarial backdrop, my own better-than-average teaching evaluations make me feel no more secure than a better-than-average Russian roulette player feels after clicking the hammer on five consecutive empty chambers. When my luck runs out, what then?

Since muttering to myself no longer packs the cathartic wallop it once did, I've gone ahead and designed an evaluation instrument of my own.

The following sample student evaluation instrument expresses both my frustration and my hope that this procedure might once again be of help to teachers. If our profession is not so lucky, at least I have fought the good fight, and with enough intellectual detachment to know that my own pet prescription probably had bugs of its own.

"Trust, but verify," a Reaganism once applied to nuclear non-proliferation efforts, is the governing philosophy of my model evaluation instrument. Given the power student evaluations have over faculty careers, it makes sense to hedge one's bets.

To be sure, student feedback is central to any meaningful evaluation of a course, but evaluators should also seek out objective evidence that can shed light on the students' subjective impressions—particularly where administrators insist on ranking professors for purposes of tenure and promotion.

If we do not rely on objective data when it is available, we cannot set the proper context for more subjective student impressions.

To put an evaluation's numerical scores into the proper perspective before they're "crunched," the interpreter of an evaluation should consult departmental and university records in order to determine objective factors before the students supply their input.

When called upon to supply such things as majors and GPAs, non-majors sometimes represent themselves as majors and C students sometimes represent themselves as A students in hopes of being taken more seriously. Accurate evaluations must eliminate such sampling errors, and effective information gathering need not compromise student anonymity vis-á-vis the professor.

Some administrators will

Responses such as 'I didn't have to work very hard' from students earning Cs or Ds should be given less weight.

doubtless respond that such an exercise is too labor intensive, but the alternative—skewed and therefore misleading evaluations—will surely cause more problems in the long run.

The first phase of the evaluation process should answer the following questions:

1) *What are the rating students' majors?*

Evaluators must know several percentages: the percentage of students who are taking the course as a major requirement, the percentage of students who are taking the course to fulfill general education requirements, and the percentage of students who are taking the course as an elective.

Getting an accurate count of majors need not compromise student anonymity, because no student's major need be correlated with his or her own responses. Totals for each of the three categories will suffice, as they will allow correlation for the whole section.

(2) *What is the aggregate cumulative GPA of the evaluating students?*

If this information is not sought, the effects of grade inflation cannot be factored into the evaluations.

Administrators need to know the GPA because they need to know how the evaluated teacher's grading compares to the grading done by other teachers in the sample group.

Once again, a student's GPA need not be correlated with his or her own responses on the evaluation form. Only the average GPA for the section need be determined.

(3) *What is the average grade in this class at the time of the evaluation?*

The students' own degree of success in the class — the average grade in the class by percentage — should also be determined to help judge the degree to which students' insights and opinions about a professor might have been prejudiced by their current grades in that teacher's course.

This information should provide the context for weighted student perceptions of a course's degree of difficulty or the amount of work required.

Numerically weighted responses supporting comments such as "I didn't have to work very hard" from students who are earning Cs or Ds should receive less consideration than other responses.

4) *What is the average of the grades assigned by the evaluated teacher's department during the current semester?*

Not only do administrators need to know which departments issue inflated grades, but all evaluated teachers have a right to know

All evaluated teachers have a right to know exactly how their grading compares with that of colleagues.

exactly how their grading compares with the grading of colleagues.

If objective standards—norms—are part of an evaluation process, is it not ethically inconsistent, as well as impractical, to hide these norms from the individuals being evaluated?

New teachers quickly learn there is a happy medium where grading is concerned: Tougher than average grading will hurt enrollments and easier than average grading will earn the enmity of colleagues.

However, intradepartmental norming need not be divisive. Faculty members need not know how others in the department rate, only how they themselves compare with the departmental average.

Moreover, this is the only factor in an evaluation instrument that demands official intradepartmental distribution. I do not need to know what the students think of other professors in my department, but I certainly do need to know exactly where I stand in relation to the departmental grade distribution average.

(5) *What is the evaluated course's format?*

Teachers of lecture courses should not be placed in the same sample group as teachers of activity-style courses in art, music performance, physical education, and applied technology.

Lectures tend to be less popular with students than is "hands-on" instruction, at least in cases where all else is equal. Not all evaluation instruments currently take this bias into consideration, and the teachers of lecture courses are often penalized accordingly.

Not only have too many designers of teaching evaluation systems tried to obtain objective information from students—non-objective sources—but they also have wrongly objectified student responses to patently subjective questions.

Many of us have seen rating forms that ask students to assign numerical scores for course organization, equity in grading, and course relevance–without demanding any supporting written response.

Numbers crunchers may consider such comments superfluous, but professors who are looking for constructive criticism can tell nothing from a number save whether the student is happy or not. Without student commentary, these professors are left to guess at what they can do to improve their courses.

Worse yet, many current evaluation formats that treat free response as "optional" also relegate it to the last page, where students are even more likely to ignore it.

Wherever student perceptions carry numerical weighting, teachers deserve written explanations.

In my own experience, the free response space is often left blank. In survey sections this happens more often than not. So how are professors to understand and respond to student concerns without some explanation for the students' judgments?

Without a written explanation for the numerical judgments, professors might also question the fitness of students to comment, without prejudice, on their grading methods or gauge the course's design and value.

Wherever student perceptions carry numerical weighting, teachers deserve written explanations. Without these explanations, the evaluative process loses its constructive character and becomes little more than a bludgeon whose use—and threatened use—can only undermine a teacher's determination to resist grade inflation.

Accordingly, students should be warned before they fill out evaluation forms that no numerical scores will count unless accompanied by written explanations. Any tabulator of evaluation data who speaks fluent English should be able to determine which comments are responsive and which are not. Therefore, the possibility for skewed results need be no greater than with strictly quantitative evaluations.

We should never be deterred from seeking non-quantifiable responses. Questions one through three below do not lend themselves neatly to numerical weighting, but they must be asked if professors are to benefit from the evaluation process.

(1) *What do you most like about this class?*

Evaluation questionnaires should not ask what the students like about their instructors. This is a "loaded" question in search of an ad hominem or ad feminem response. If students genuinely liked something about the instructor a great deal, they will include a remark to that effect here.

(2) *What do you least like about this class?*

As in the previous question, the evaluation instrument should make no specific reference to the instructor. If the evaluating student encountered a significant problem with the instructor's personality, teaching, or grading, it will surface here anyway.

Conversely, if the questionnaire specifically seeks criticism of the instructor, a higher percentage of the complaints will be along the lines of "wears weird ties." Even students who are satisfied with the instruction may feel compelled to complain about something.

(3) *What can the teacher do to improve this class?*

Questionnaire writers should avoid such phrases as 'rapport with students' and 'relate to students.'

This is probably the most important question that can be asked, but, because it doesn't lend itself to numerical weighting, it is asked ever less frequently.

Fortunately, the following questions can be numerically weighted. This should appease the numbers crunchers who have trusted numerical values for too long without demanding contextual verification.

A scale of one to seven (maximum), with four being average, should suffice. But another warning is in order: Those who do not think an evaluation instrument worthwhile unless it ranks faculty members against one another should consider the potentially divisive effects that such competition can have on a department.

Doubtless some administrators who employ divide-and-conquer management strategies hope that precisely this will happen, but those who act in good faith should recognize that the healthiest kind of competition does not depend on a peer's failure.

Using this approach, faculty members need only to focus on besting their own scores from the previous semester, rather than having to obsess like a perennially frustrated Cubs fan about their place in the standings.

The questions:

(4) *How well does the professor get ideas across to students?*

Questionnaire writers should avoid such phrases as "rapport with students" and "relate to students" as these are fraught with misleading connotations.

Specifically, teachers should not have to be "cool" or be able to relate to students on all levels in order express a specific idea about the course material, and students should not expect this.

(5) *Compared to other faculty members at this university, how enthusiastic was the professor?*

This question may well be the most problematic of all if phrased any other way. Too often, students are likely to equate "enthusiastic" with "entertaining." Many faculty members already suspect that they are unfairly expected to deliver what television, movies, and stand up comics deliver, and their concerns are often justified.

By posing separate questions about narrower aspects, enthusiasm and interest, and stressing that the sample group is teachers only, the author of an evaluation instrument lessens—but admittedly cannot entirely overcome—the possibility of a misunderstanding.

(6) *Compared to other faculty members at this university, how interesting was the professor?*

Students need to be reminded that the sample group is composed solely of academics at their university.

Like the previous question, this one can be easily misconstrued, but the possibility for such a mistake is lessened if interest level is treated separately from enthusiasm.

Again, students need to be reminded that the sample group is composed solely of academics at their university.

(7) *Compared to other teachers at this university, how competent was the professor?*

Once more, by emphasizing to the students that their professor is to be compared only to others in the sample group, and not to some idealized, inaccurate image the evaluation instrument will serve the rated faculty members more reliably.

Where numerous careers are under scrutiny, abstract ideals shouldn't be substituted for realistic performance expectations and goals.

(8) *Compared to other course material at this university, how interesting was the course material?*

This is an oblique but often reliable way of determining if the rated faculty member is "getting through" to the student. However, for survey sections, where the overwhelming majority of students are non-majors, the percentage of non-majors must be factored in.

Equally important, the instructor of a survey section who manages to keep a high percentage of general education students interested merits at least some additional credit.

(9) *Compared to other courses at this university, how difficult was this course?*

Here, too, the percentage of non-majors should be factored in. Moreover, values for perceived course difficulty should be noted in conjunction with the rating students' performance in the course at the time of the evaluation.

As noted, students who found it easy to earn a "C" in the course and were not trying for anything better will skew the evaluation's results if such a qualifying factor is overlooked.

(10) *Compared to other courses at this university, how much work was required?*

The same qualifications pertinent to question 9 must be applied here as well.

(11) *Compared to your experiences in other courses at this university, how much did you learn?*

The student's major and performance in the course are crucial elements here and must be factored in. A required written response in which rating students cite examples would be especially useful.

(12) *If you needed to take another course in this department, how likely would you be to take it from this professor?*

The written response to this

It is unfair to perpetuate a system that benefits students and administrators while threatening faculty members.

question will include any positive or negative ad hominem or ad feminem remarks not noted in questions one and two.

(13) *If you needed an elective that could be satisfied with any course offered at this university, how likely would you be to take another course from this department?*

Once students have given their impressions by assigning numerical scores and commenting on them, the subjective aspect of the evaluation can be considered in light of grade point averages and current performances of those students.

The mathematical formulae used here may require fine tuning, but, as student impressions will no longer be considered out of context, this system is an improvement over any other norming scheme I have yet seen.

Once we make clear the context of the student opinions we solicit, teacher evaluation instruments will serve teachers as well as students and administrators. Conversely, it is unfair to perpetuate a system that benefits students and administrators while threatening faculty members.

(14) *Estimated Course Impact.*

The average of scores for the student responses to questions four through 13 should be divided by the student performance level indicator — the percentage sought in question three.

Once the estimated course impact scores in the sample group are determined, the administration can, if it must, compare the rated professor to his colleagues.

(15) *Estimated Course Difficulty.*

The average score for questions nine through 11 should also be divided by the average student performance percentage from question three.

As with estimated course impact, the rated professor can be assigned a percentile once the data from his or her sample group have been tabulated.

The above suggestions are the potential components of an improved evaluation system, not a perfect one. Any formula for norming, whether it purports to be scientific or not, can be used justly or unjustly. The ultimate impact hinges on the intentions of the users.

The instrument I describe in this article will give administrators a way to rank faculty members and students the voice they deserve. Because this instrument demands written responses from students, it will also provide instructors with what *they* need: useful information that can help them improve instruction. ■