References

Williams, W., & Ceci, S. (1997). "How'm I Doing?" Problems with Student Ratings of Instructors and Courses. *Change*, *29*(5), 12. Retrieved Friday, April 13, 2007 from the ERIC database.

"HOW' M I DOING?"

Problems With Student Ratings of Instructors and Courses

When Ed Koch was mayor of New York City, every time he appeared in public, his tried-and-true motto was to ask every person, crowd, newspaper and TV reporter, and even leashed pet,

"How'm I doing?" It was Koch's way of showing that he cared about the public's opinion of his performance and that, for an elected official, there's no such thing as being too concerned about your constituents' opinions.

College and university faculty might feel they have little in common with Mayor Koch, but whether they realize it or not, they are wrong. Just as the New York City voters passed judgment on the mayor, so, too, do crowds of students silently pass judgment on their professors. But the students don't use a voting booth: their judgment vehicle--the course evaluation form--is handed to them by the instructor on the last day of class. Set loose with the freedom of anonymity, students use numerical scales to rate teachers for their "organization," "openness to alternative points of view," "availability outside of class," "knowledge of the course material," and so on. Students also rate the course as a whole on dimensions such as its overall quality, fairness of exams, and its interest level.

Then what happens? At one time--when today's professors were college students--the course evaluations were read by the teacher as a form of performance feedback for improving the course. Reappraisal of policies, techniques, and course materials was thought to be usefully guided by such feedback. But in the 1960s, student groups got hold of the course evaluations, and on many campuses they began publishing course guides peppered with biting comments. When one of us was a student at Columbia 20 years ago, the course guide contained comments like "Physics 110, a.k.a. Physics for Poets, unfortunately taught by neither." Some teachers were hung out to dry once the course evaluations became public; others enjoyed enormous enrollments and overnight celebrity.

Not only were course evaluations circulated to student groups, but--even more nerve-wracking for the untenured professor--they began to be circulated to department chairs, deans, and voting faculty in general. At this point, course evaluations took on increasing importance in the lives of teachers. A professor of organic chemistry, for example, might be forced to explain his lower-than-average ratings by pointing to the rigor and monotony inherent in learning organic chemistry, while a modern film teacher might bask in uniformly excellent ratings after showing five feature films over the semester. An instructor might even witness enormous fluctuations in ratings for the same course taught at different points in time, as reported by A. G. Greenwald in his award address at the 1995 American Psychological Association annual meeting. (See Resources.)

Today the picture is even more worrisome. When an assistant professor comes up for tenure, her or his course evaluations are amassed for every course taught over the preceding five years, average ratings are computed, and these ratings are then compared to department-wide and institutionwide averages for similar courses at similar levels. Deans and tenure review committees take these comparisons very seriously. (Coincidentally, while writing this article, one of us heard from a friend elsewhere who was just denied tenure by a university review committee on the heels of a unanimously positive departmental vote. The reason? "Below-average" course evaluations.) Even after being tenured, professors often find that their yearly pay increases and future promotions are tied to course evaluations.

Why have such evaluations assumed greater importance in the past 10 to 20 years? One reason is a heightened desire for objective grounds on which to evaluate faculty for reappointment, tenure, and promotion decisions; on the surface, at least, student evaluations produce numbers, which seem not to lie. Moreover, it is argued, such evaluations are better placed in the hands of students than of colleagues, since the latter could have ulterior motives. Indeed, student evaluations seem to fulfill demands for greater accountability on the part of faculty. And, as Greenwald points out, they are cheaper and easier to collect. Consequently, at most institutions, the use of student ratings has superseded peer visitation and evaluations of syllabi.

Given the omnipresent power of course evaluations today, one might think that their accuracy and appropriateness had been extensively studied. Although there have been many studies examining the internal consistency and other statistical properties of ratings (for reviews, see Greenwald; Marsh and Duncan, in Resources), research on the basic fairness of these ratings is lacking. In a review of the literature on the evaluation of teaching effectiveness, one is immediately impressed (or depressed, depending on one's perspective) by the absence of experimental data to validate such measures. (Greenwald shows that

experimental research on teacher ratings, conducted primarily during the 1970s, led to a different conclusion from that reached in the correlational analyses of the 1980s, but, because of ethical prohibitions, the experimental evidence ceased to be produced: "Those experiments are difficult to repeat in the 1990s, because their grade manipulations imposed stresses and used deceptions that university human subjects review committees do not look kindly upon.")

Today's "validation" of student ratings of teaching effectiveness comes in the form of aggregate statistical summaries demonstrating that such ratings are quite reliable; do not shift with the age of the student; are related to the student's grade in the course; and are not the result of simple affective or physical qualities of the instructor such as attractiveness, gender, or age, but are associated with presumably more valid qualities such as "warmth," "supportiveness," "dominance," and "confidence." (See Abrami, d'Apollonia, and Cohen; Erdle, Murray, and Rushton; Feldman; and Marsh, in Resources.)

Alongside the absence of experimental inquiry into the factors that might influence students' judgments of instructors' teaching ability is a handful of fascinating studies into the micro-level processes that students engage in when they are asked to rate teachers. This research shows that students often arrive at summative judgments about their teachers very quickly, even before they have a chance to learn about the teacher's knowledge, fairness, organization, or grading policy.

In the most provocative of these studies, Ambady and Rosenthal have shown that students arrive at opinions about teachers within seconds of being exposed to them. (See Resources.) In one experiment, the authors showed Harvard students video clips (without audio) of graduate teaching fellows as they lectured. Because there was no sound, the students had only 30 seconds of content-free video to watch. Ambady and Rosenthal reported that the average end-of-semester student ratings of these graduate teachers correlated with the mean of 15 ratings made by the students who watched only 30 seconds of content-free video at an impressive .76 level.

In effect, based on video clips just half a minute long, complete strangers were able to predict quite accurately the ratings of teachers by students who had interacted with these teachers over the course of a whole semester! Furthermore, these predictions retained their accuracy after the researchers adjusted them for physical appearance of the teachers, indicating that the judges were picking up very subtle nonverbal cues. In subsequent experiments, these researchers found similarly high correlations (.68) between ratings generated by students who watched even briefer content-free video of high school teachers (as little as three two-second segments) and ratings generated by the teachers' high school

principals.

It was against the twin backdrops of 1) greater use of student ratings in pay, promotion, and tenure decisions, and 2) demonstrations that student ratings can be predicted even in the absence of substantive information about the teacher, that we conducted the present study. We sought to provide what we regard as a missing link in these backdrops: an analysis of a direct link between student evaluations and what Ambady and Rosenthal term a teacher's "sending accuracy."

One absence in the empirical research on teaching effectiveness is the experimental disentanglement of variables that may be collinear in the real world. For instance, teachers judged "warm" or "confident" may in fact possess greater knowledge, organization, and actual teaching skill than do their peers who are judged to be less so. Only an experimental mapping of factors such as knowledge level and organizational clarity across "sending accuracy" in variables such as warmth and confidence can inform the point. Given the professional responsibility of every instructor to provide the highest quality educational experience he or she is capable of providing, only the rather unusual circumstances of an "experiment of nature" (see Bronfenbrenner, in Resources) permit such an experimental mapping without violating ethical strictures.

What is known about factors that might influence student ratings of teaching effectiveness? Do such ratings measure what we assume they measure? Are student ratings related to other indicia of teaching effectiveness (such as amount learned)? Are students' evaluations of teachers fair and accurate--do the marks students give their teachers reflect meaningful aspects of their teachers' performance? Do students grade fairly the instructor's teaching ability and knowledge of course material? These are the types of questions we sought to answer.

Our goal was to examine the degree to which students' evaluations of instructors vary due to changes in instructor behavior that do not affect a course's information content and learning value. Therefore, our main question was, What specific mechanisms are involved in student evaluations? Can we elevate student evaluations of teaching effectiveness if we change one "content-free" (or, stylistic) factor but hold all content-relevant factors constant?

BACKGROUND OF THE PRESENT STUDY

Many professors suspect that their student ratings fluctuate systematically in accordance with factors that have little to do with the actual content of their lectures, reading materials, or grading policies. In his 1995 award address at

APA, Greenwald reported that the student ratings of his 1989 honors seminar in social psychology placed him in the top 10 percent of faculty et the University of Washington. The following year, however, his ratings for the same course plummeted: "Having taught the same course in 1990 according to the same plan...imagine my surprise when my 1990 ratings turned out to be the lowest I had yet received, placing me in the second-lowest decile of the university's faculty....Did I think, Wow, what on earth did I do I wrong? Yes."

Similar expressions of wonderment can be heard throughout the academy, particularly among new faculty. To date, we know that grading leniency tends to increase student ratings, but we do not know what other factors do. In fact, the dominant view is that student ratings are the "best available" means for assessing a lecturer's competence: "In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation." (See Cashin, in Resources.)

But are they? The present study grew out of an unexpected naturalistic experiment. One of us (SJC) had been teaching an undergraduate course in developmental psychology since 1977, when his university invited him to participate in a teaching skills workshop. Did he need a teaching skills workshop? His student evaluations for the course were roughly average for courses at the same level (introductory/survey) but, as the university memo said, "there is always room for improvement."

But how much room for improvement? We decided to take advantage of the opportunity provided by the teaching skills workshop to answer this question in a controlled manner. First, we collected the usual student ratings for the fall semester before the teaching skills intervention. Next, over the inter-session the professor (SJC) attended the teaching skills workshop. During the spring semester following the workshop, he attempted to teach the identical course he had taught in the fall, using the identical materials on a lecture-by-lecture basis, with one exception: his presentation style was more enthusiastic. Specifically, he spoke with more pitch variability and used more gestures while lecturing, the "enthusiastic style" recommended in the teaching skills workshop. Given this single change, we were presented with a natural occasion to assess the impact on student evaluations of a purely stylistic--as opposed to content-based-change in the course.

METHOD Subjects

A total of 472 students enrolled in the course "Developmental Psychology" at Cornell University participated in this study--243 during the fall semester and 229 during the spring. This introductory survey course has been taught since

1977 by the second author.

The fall and spring groups were similar with respect to those demographic characteristics for which data were available. Specifically, the two class groups were comprised of similar percentages of freshmen (55 percent versus 53 percent), sophomores (21 percent versus 21 percent), juniors (14 percent versus 15 percent), and seniors (8 percent versus 9 percent); they were similar in gender composition (64 percent females versus 60 percent females), and in the proportions of students in each major-subject area (social science: 45 percent versus 46 percent; humanities: 21 percent versus 17 percent; physics/mathlengineering: 20 percent versus 20 percent; art/architecture/planning: 9 percent versus 12 percent; agriculture and life science: 6 percent versus 4 percent).

Procedure

Description of teaching skills workshop. The teaching skills workshop attended by the instructor was taught by a professional media consultant who was neither a psychologist nor an academic. The workshop was completely free of academic content related to teaching in a particular discipline. Faculty from several departments took part in the training, which consisted of three sessions. During the training, the media consultant explained how to improve one's presentation style while lecturing. Participants were videotaped giving a sample lecture; the consultant critiqued their presentation styles and offered suggestions for improvement. Specifically, the media consultant suggested that instructors vary their voice pitch and use hand gestures in an attempt to be perceived as more enthusiastic.

The workshop did not cover content-related improvements that instructors might use to improve their teaching. In the case of the present course, the instructor had spent the years since 1977 refining its content, format, examinations, grading policies, and other content-related aspects. The area open to improvement seemed to lie in the instructor's presentation style, and the teaching-skills workshop had shown how that might be done.

Design of study. A conscious effort was made to teach the "Developmental Psychology" course exactly the same way in the spring as it had been taught in the fall. The only systematic difference between the two semesters was the professor's use in the spring of a more enthusiastic presentation style, whereas he had taught in his customary manner during the fall semester (before the teaching skills workshop). With the exception of this change in presentation style, all other aspects of the course were as identical as possible in the two semesters.

Specifically: 1) the same syllabus, textbook, and reserve readings were used each semester; 2) the same overhead transparencies were used at the same relative points each semester; 3) the same teaching aids (slides, videos, demonstrations) were used at the same points in each semester; 4) the identical exams and quizzes were given in each semester (with appropriate safeguards against their being removed from testing rooms by students); 5) nearly identical lectures were given in both semesters; 6) the course met on the same days of the week, at the same time, and in the same room both semesters; and, finally, 7) the course had the same ratio of teaching assistants to students each semester.

The point about the lectures being nearly identical in both semesters requires amplification. During the first semester, each lecture was audio recorded to preserve its contents (as was usually done by the professor for the benefit of students who miss class). Before the professor gave the same lecture in spring, he attempted to memorize what he had said in that lecture the prior fall. This was not as difficult as it might seem, for two reasons. First, the professor has been teaching this course since 1977, and on several occasions has taught it two to three times per year, thus resulting in a highly rehearsed set of lectures. Second, each lecture is accompanied by overhead transparencies that provide a detailed outline for that lecture. It is difficult to avoid repeating all of the substantive points made during the previous semester, inasmuch as the professor and class follow these outlines point by point.

Materials

In the final week of the course, students were asked to rate the course and the instructor by responding to 10 questions on a 1-to-5 scale. In order to avoid response bias, half of the students in each semester received a form with the scale anchored as 1 = Very Much/Excellent and 5 = Very Little/Very Poor, while the other half of the students in each semester received a form anchored in the opposite direction (with a 1 = Very Little/Very Poor, etc.). Additionally, students were asked questions about their major, class year, gender, and how well they were doing in the course (number of points earned). Finally, they were asked whether they would recommend the course to a friend (yes, no, or unsure). These evaluations were anonymous, and were conducted in the students' small-group discussion sections supervised by a teaching assistant. Students were given approximately 20 minutes to complete the evaluation form. Because the ratings were administered during class meeting times, there were very few missing data for either semester.

VALIDATING FINDINGS

Before turning to findings, we provide here some preliminary validity checks on

the success of the professor's attempt to teach more "enthusiastically," and discuss a caveat regarding the interpretation and generalizability of our findings.

Because our analysis rests on the assumption that the students' ratings were accurate reflections of what they believed, we sought some indication that they took their ratings seriously and reported accurately. Toward this end, we verified students' reports of their progress in the course by checking the accuracy of their claims about how many points they had earned on tests and quizzes.

We did this by taking the ratings of 62 students for whom we had determinative identifying information (that is, we were able to identify these students by virtue of their answers to questions about the identity of their teaching assistant, their academic major, and their matriculating class). We did this after the semester was completed and by asking an assistant who was blind to the study's hypotheses to verify these students' point estimations. Of these 62 students, 57 gave point estimations that were identical to the ones that were officially recorded. Thus, although students assumed that their ratings were confidential (and they were), they nevertheless were quite accurate in their estimates of how well they were doing in the course. Absent evidence to the contrary, we take this result to indicate that students took the rating task seriously and that their ratings reflect accurate reporting.

Second, because the central question in this study concerned the difference made by the professor's presentation style in how students judged his other qualities as a teacher, it was desirable to document the fact of this difference. One item on the Cornell Student Evaluation Form asked directly, "How enthusiastic was the professor?" Obviously, this question was directly relevant to the instructor's change in presentation style. In the fall semester, students rated his enthusiasm level 2.14 (SD = .94) out of a possible 5, where 1 reflects "Very Little/Very Poor" and 5 reflects "Very Much/Excellent." (We corrected the half of the forms that used the reversed scale to avoid response bias.) In the spring term (after the teaching-skills training), students' mean rating of the professor's enthusiasm was 4.21 (SD = .83). So, the professor was indeed rated as quite enthused during the spring, when he tried to teach with a more enthusiastic style. This difference in enthusiasm level was highly reliable, $p < .0001$.

Third, an analysis of the content of the fall and spring semesters' lectures was conducted by two independent raters blind to the research hypotheses. These individuals listened to and compared eight randomly selected matched pairs of lecture tapes from the course (such as the lecture on Piaget's theory from both fall and spring). This analysis revealed that there was 100 percent (perfect) overlap in idea units covered in class (Cohen's Kappa = 1.0). Every idea mentioned in the fall semester was also mentioned in spring, and there were no extra ideas

mentioned in either semester. Furthermore, there was a very high degree of actual verbatim overlap between the two semesters, a result of the professor's successful attempt to memorize his prior semester's lecture before giving the corresponding second-semester lecture.

From these three validity checks we are able to demonstrate that students took the evaluation seriously, that there was a reliable difference in the professor's demeanor between the two semesters, and that material virtually identical in content was taught both times (as well as, in addition, identical textbooks, exams, and teaching aids). To further rule out other factors that could be invoked to account for differences in students' evaluations unrelated to our main hypothesis, we established that the composition of the class was the same both semesters, as was the gender breakdown of the class and the year of matriculation (an almost perfect surrogate for student age, since Cornell is not a commuter campus and has very few "mature students" or students who do not finish in four years). Finally, we established that course ratings at Cornell do not differ by semester: fall and spring summative ratings over a three-year period are nearly identical, so whatever differences emerged in this study cannot be attributed to seasonal variations in ratings.

This brings us to a caveat regarding the interpretation and generalizability of our findings. By its nature, this type of research precludes the use of double-blind controls and random assignment typical in true experiments. There is no ethical way for researchers to deliberately manipulate variables related to teaching effectiveness in order to assess their effects on student performance; it will always be inappropriate for an instructor to provide a less-high-quality learning environment than he or she is capable of providing at a given point in time for the purpose of conducting an experiment. However, any instructor is entitled, and in fact is expected, to improve her or his teaching in any way possible over time.

To the extent that an instructor's attempt to improve is focused on purely stylistic aspects of presentation, and does not affect course content, we have the basis for a naturalistic experiment with objectively verifiable measures and controls. In this research, we took numerous precautions to ensure that such controls were in place. However, we acknowledge that the experimenter (professor) was not blind to the research hypothesis--nor for ethical reasons could he be--and we acknowledge that this fact could in principle affect the interpretation and generalizability of our findings. Given that the content analysis of lectures from the two semesters showed that the same content was covered nearly verbatim at the same point in the semester by a veteran instructor, and given that the same tests were used, and that in every other aspect the two courses were identical, we are unable to think of any factor that could be responsible for changes in student

ratings from fall to spring except for the instructor's presentation style.

With our validations complete, and bearing in mind our caveat, we next examined the effect on student ratings prompted by the professor's change in presentation style.

Ratings of the Instructor

The means and standard deviations for each item rating the instructor and the course appear in Table 1, while Chart 2 is a histogram portraying the results. The scale used was anchored as follows: 1 = Very Little/Very Poor, 2 = A Little/Poor, 3 = Average, 4 = A Lot/Good, and 5 = Very Much/Excellent. (As already noted, we "reflected" the ratings for the one-half of the students whose scales went from a 5 to a 1 instead of from a 1 to a 5, so that all ratings were on a scale where 1 was the lowest rating and 5 was the highest.)

The first question was, "How knowledgeable is the instructor?" The mean rating was 3.61 for the fall semester, and 4.05 for spring. The difference was highly significant, $t(470) = 5.33$, $p < .0001$. What this means is that the instructor was rated as substantially more knowledgeable by the students when he used a more enthusiastic presentation style. This striking pattern of results held for all of the instructor-related variables: all comparisons were significant at the $p < .0001$ level.

The second question was, "How tolerant of others' points of view is the instructor?" The means here were 2.55 (fall) and 3.48 (spring), meaning that the instructor was rated as far more tolerant of students' viewpoints when he used a more enthusiastic presentation style ($t = 11.80$, $p < .0001$). The third question was, "How enthusiastic is the instructor?" For this question, the means were 2.14 (fall) and 4.21 (spring)--over a two-standard-deviation increase ($t = 25.58$, $p < .0001$)! The fourth question was, "How accessible is the instructor outside of class time?" The means were 2.99 (fall) and 4.06 (spring), showing that the more enthusiastic instructor was deemed far more accessible by his students ($t = 13.57$, $p < .0001$)--a finding that certainly does not reflect reality because the instructor maintained identical office hours both semesters and made himself equally available for student appointments and consultations over both semesters. The final instructor-related question was, "How organized is the instructor?" Mean responses were 3.18 (fall) and 4.09 (spring) ($t = 11.07$, $p < .0001$). Again, the perception of the instructor's level of organization was strongly affected by his presentation style, even though his organization was the same both semesters (that is, he used the identical syllabus and overheads and his lectures were nearly identical).

Finally, a summary rating representing the mean of all instructor ratings moved from a 3.08 (fall) to a 3.92 (spring), well over a standard deviation increase ($t = 14.04, p < .0001$). In sum, by working to increase the enthusiasm of his presentations during the spring, the instructor obtained higher student ratings not only of his enthusiasm level (as expected) but also of several characteristics that were held constant over the two semesters (his knowledge, tolerance, degree of organization, and accessibility).

Ratings of the Course

The ratings of the course followed the same pattern as did the ratings of the instructor. The first question was, "How much did you learn in this course?" The means were 2.93 (fall) and 4.05 (spring) ($t = 12.82, p < .0001$). The students in spring may have thought they learned more, but in fact, they had not: the end-of-semester point totals for the identical sets of exams (based on nearly identical lectures and identical texts) were virtually identical for the two semesters in both mean and standard deviation. The fact that the students say they learned markedly different amounts of information--from less than "average" to more than "a lot" on the rating scale--in the presence of differences in the professor's style is staggering. Perhaps more shocking is the fact that the students were wrong about how much they had learned, as reflected in the objective tests and final grades given in the two semesters, which were virtually identical!

The second question was, "How clearly stated are the course expectations, grading policy, and goals?" Here the means were 3.22 (fall) and 4.00 (spring) ($t = 9.03$, $p < .0001$)--for the identical syllabus, quizzes, exams, and procedures. Again, the course's constant goals and grading policy seemed ever so much better when the instructor used a more enthusiastic style. The third question was, "How fair is the grading system used?" The mean responses were 3.03 (fall) and 3.72 (spring) ($t = 7.95$, $p < .0001$), showing that even for an objective criterion, which in fact was held absolutely constant across the two semesters, the students perceived a substantial difference. (Note that the syllabus described in detail the grading system and method for determining final grades, which was identical over the two semesters.)

The fourth question was, "How would you rate the text for this course?" Note, again, that the same exact text (same edition) was used both semesters, yet the ratings varied widely: 2.06 (fall) and 2.98 (spring) ($t = 12.51$, $p < .0001$). The same text that was rated "poor" in the fall became "average" in spring, when the instructor spoke more enthusiastically.

Finally, the fifth--and in the eyes of the institution, the most important--question asked for the student's summary rating: "Among courses that you have taken at

this level, how would you rate this one?" The mean responses were 2.50 (halfway between "poor" and "average" for fall) and 3.91 ("good" for spring) ($t = 16.50$, $p < .0001$)--a difference of more than 1.5 standard deviation units. (The reader should note that although comparing ratings across instructors is risky because of the many differences in how they teach the same course--such as number of lectures per week; number and type of tests, assignments, and projects; existence of study sections--in this study the instructor's mean fall overall course rating of 2.50 fell between the mean course ratings for the previous and subsequent instructors of the same course.) The overall course rating thus showed extreme modifiability due to an instructor's presentation style.

A summary course rating representing the mean of the five individual course ratings was also computed. Its means were 2.70 (fall) and 3.65 (spring) ($t = 14.79$, $p < .0001$), confirming the substantial change in ratings of the course due to instructor's presentation style. In addition, students were asked whether they would recommend the course to a friend (no = 1, unsure = 2, or yes = 3). The mean ratings for this question were 2.36 (fall) and 2.81 (spring) ($t = 7.49$, $p < .0001$). While students in the spring semester were nearly unanimous in stating they would recommend the course, many students in the fall semester were unsure about recommending the course.

Additional Analyses

As mentioned above, we compared the final course performance for students in the two semesters: their quiz and test points earned, and their final grades as well. Results showed nearly identical performance by students in both semesters. This finding reflects that students in fact learned the same amount during both semesters as measured by objective assessments. This is not surprising inasmuch as the courses were virtually identical and were taught by a seasoned teacher with many years' experience in teaching the same course. (It is always possible, however, that real differences in learning occurred that were too subtle to be detected by the battery of quizzes and exams given.)

We also looked at the degree to which students' performance in the course influenced their ratings of the instructor and the course. As expected on the basis of past studies by Abrami et al., Feldman, and Greenwald, there was a strong positive correlation between grade received and overall course rating ($N = 468$, $r = .42$, $p < .0001$). This value is close to the "grand mean" of .40 reported in the largest meta-analysis done to date by Abrami et al. (Grade received also predicted all of the other instructor and course ratings, a common halo-type effect. See Greenwald for a review.)

We checked, too, for gender effects in student ratings and in performance in the

course; we found no gender effects in these data. We also computed the intercorrelations for the set of questions rating the instructor as well as for the set of questions rating the course. Instructor-rating intercorrelations ranged from .37 to .68 with a mean of .53. Course-rating intercorrelations ranged from .50 to .68 with a mean of .56. These results show a moderate to high degree of interrelationship for the student ratings.

DISCUSSION

Our point is not especially that content-free stylistic changes can cause students to like a course more or less; nor is it that students' general affect toward a course influences their ratings of multiple aspects of the course and its instructor (halo effects). What is most meaningful about our results is the magnitude of the changes in students' evaluations due to a content-free stylistic change by the instructor, and the challenge this poses to widespread assumptions about the validity of student ratings.

Our results also show that the substantial changes in student ratings we report were not associated with changes in the amount students learned. The substantial improvement in spring-semester ratings was not due to having a more knowledgeable instructor, better materials and teaching aids, a fairer grading policy, better organization, and so on: the increases occurred because the instructor used a more enthusiastic teaching style. With coaching, teaching in a more enthusiastic style is a fairly easy change to effect. The astonishing thing is that so simple a change can make the difference between being awarded tenure or not, or in receiving other consequential rewards (teaching awards, merit pay increments, and so on).

In light of our findings, then, it is disconcerting to go back to Cashin's assertion that "in general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation." This prevailing sentiment about the validity of student ratings led Greenwald to lament that "for the past 15 years well-respected researchers have asserted that it is acceptable to treat student ratings as construct-valid measures of instructional quality."

What do our results imply about the way we do business in the academy, and specifically about the way we evaluate teaching effectiveness? One would like to believe that psychologists follow their own earnest advice about the potential misuses of evaluation (see AERA et al., in Resources), and that we would never advocate the use of any device unless it possesses satisfactory validation data.

The sad truth is that we psychologists know very little as yet about how students

arrive at their judgments about teaching effectiveness. Based on the present data, we know that it is at least possible for student ratings to be extremely systematic, and reliable, yet invalid! In the present example, ratings of the identical text, grading policy, and so on, were "colored" by students' reactions to a content-free change in presentation style.

Whatever significance one attaches to these findings, our own belief is that their main usefulness is as an "existence proof." They vividly demonstrate what some teachers have long believed but were unable to document empirically, namely, that factors unrelated to actual teaching effectiveness (such as variation in a professor's voice) can exert a sizable influence on student ratings of that same professor's knowledge, organization, and basic fairness.

We do not maintain that the effects of presentation style we found are the strongest we could have obtained for this or other sources of variation (future research must examine other variables, especially micro-level behaviors), or even that our findings were inevitable. Indeed, our findings are limited by their case-study nature, coming as they do from a single professor teaching a single course within a single university.

Nor do we claim that these results obviate the many valid uses for student ratings (for example, if 60 percent of students independently rate a professor as "persistently tardy," or the exams as "not reflecting the material taught in class lectures," there is undoubtedly a high degree of validity to their claims).

The present findings do suggest that in searching for better and fairer means of evaluating teaching effectiveness and providing better bases for reappraisal of one's teaching, we need to experiment with alternative methods of soliciting students' opinions. Perhaps framing questions more in terms of concrete activities ("Does the professor show up on time?") and course performance (total number of points earned, length of retention of material) will prove superior to asking about general factors that become permeated with impressions unrelated to whatever it is we assume we are assessing.

Whatever course of action we decide to take, one practical thing is clear from these results: teaching faculty should be given the opportunity to train in techniques that can enhance their student ratings, especially if such ratings are to be used by administrators in recommendations for tenure and promotion. The present results provide some admittedly limited empirical support for the value and importance of such training. (See Woolfolk and Brooks, in Resources.)

Given all of the limitations of a naturalistic case study, our modest study nevertheless shows that student ratings are far from the bias-free indicators of

instructor effectiveness that many have touted them to be. Moreover, student ratings can make or break the careers of instructors on grounds unrelated to objective measures of student learning, and for factors correctable with minor coaching. Our findings do not imply that we should abandon the use of student ratings--only our uncritical acceptance of their validity. (Greenwald notes that even if student ratings measure only the popularity of an instructor, independent of teaching effectiveness, this measure of student affect may still predict a student's willingness to register for future course work from that instructor, or even in that discipline.)

Which brings us back to Mayor Ed Koch and his infamous question, "How'm I doing?" Koch knew what many academics are just now learning--that it pays to be attentive to one's constituents, be they voters or students. Political leaders know that it is not only what they do and how well they do it that matters, but more importantly, how they appear while doing it. Today, all instructors would be well advised to ask their students frequently, "How'm I doing?"--and listen carefully to the answer. As in politics, however, the answer may have more to do with style than substance.

Author note: The authors gratefully acknowledge the generous assistance of Jason Millman of Cornell University, Wayne Camara of The College Board, and Peter Salovey of Yale University.

TEACHERS: "WHO WE ARE MATTERS"

The following is excerpted from the article "The `Who' of Teaching," in the April 16, 1997 issue of Education Week. Its authors, lames M. Banner Jr. and Harold C. Cannon, wrote these words before seeing the Williams and Ceci article. Banner and Cannon are the authors of The Elements of Teaching, published this past April by Yale University Press.

....We recall the great teachers of our lives principally as characters--for the stories they told, the distinctive ways they kept order, their extraordinary hold on our attention, their gravitas, or their mannerisms and expressions--rather than for what they knew or how they taught us, which we are likely to have forgotten. These teachers seemed great as human beings before we knew them as superb scholars or ingenious instructors....

The truth is that who we are matters to our teaching every bit as much as what we teach and how we choose to teach it. In fact, our characters and personalities determine the quality and effectiveness of teaching long before what we know and how we present it even come into play. Questions about their teachers, especially about their personal qualities, crowd into our students' minds before

they're conscious of the fact: Who is this person? Do I like (or dislike) her? What do I like (or dislike) about him? How can I find out more about her? It is the qualities of our selves and characters that are immediately on display when we try to instruct other people, whether they be kindergarten pupils, graduate students, our own children, or employees and colleagues, and it is these qualities, as much as our knowledge and techniques, that are likely to count in determining our effectiveness....(Reprinted with permission from Education Week, Vol. 16, No. 29, April 16,1997.)

The following is excerpted from the article "A Critique of the Five Points Approach," by Laura Border, which appeared in The National Teaching & Learning Forum, Vol. 6, No. 3. Laura Border is director of the Graduate Teacher Program at the University of Colorado at Boulder.

....Show enthusiasm! we are told. Such thoughtless exhortation makes my blood boil. And...I am not alone in my fury!

Shouldn't those striving to improve teaching be a bit more thoughtful? What is enthusiasm anyway? How is it demonstrated? Why do students perceive some teachers as enthusiastic and others as feckless? I have heard lots of faculty say, "Oh hogwash, if they want me to be a cheerleader, forget it...."

I don't blame faculty for expressing such frustration.

But if we take seriously the notion that teaching can be improved and if we want faculty to take it seriously, we are going to have to dig deeper into concepts like enthusiasm than we have done up until now. We need to search for visible and obvious teaching behaviors that lead students to judge faculty as enthusiastic. And given that, since personalities and teaching styles vary, some behaviors look insincere on some people and full of conviction on others, we can see that even a supposedly simple matter like showing enthusiasm can be complex and difficult.

Because "enthusiasm" is often misread as "cheerleading," many of us shy away from it. But let's think about what we liked when we were students. One of the most riveting teachers I ever encountered was physically plain, never moved, and spoke almost in a whisper. Yet you could have heard a pin drop in her class when she was talking. No one missed class. Listening to her was simply fascinating. She cared about her work and our learning. She was enthusiastic....(Used with permission of James Rhem & Associates, Inc. and The Oryx Press, 4041 N. Central Ave., Suite 700, Phoenix, AZ, 85012. (800) 279-6799. @) 1997. Annual subscription prices: $39 for print, $32 online at http://www.ntlf.com.)

RESOURCES

Abrami, P., S. d'Apollonia, and P. Cohen. "Validity of Student Ratings of Instruction: a What We Know and What We Do Not," Journal of Educational Psychology, Vol. 82, 1990, pp. 219-231.

Ambady, N. and R. Rosenthal. "Half a Minute: Predicting Teacher Evaluations From Thin Slices of Nonverbal Behavior and Physical Attractiveness," Journal of Personality and Social Psychology, Vol. 64, 1993, pp. 431-441.

Ambady, N. and R. Rosenthal. "Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-analysis," Psychological Bulletin, Vol. 111, 1992, pp. 256-274.

American Educational Research Association, American Psychological Association, and National Council of Measurement in Education. Standards for Educational and Psychological Testing, Washington, DC: American Psychological Association Books, 1985.

Bronfenbrenner, U. "Toward an Experimental Ecology of Human Development," American Psychologist, Vol. 32, 1977, pp. 513-531.

Cashin, W. E. "Student Ratings of Teaching: IDEA Paper No. 32," Manhatten, KS: Kansas State University, Center for Faculty Evaluation and Development, 1995.

Erdle, S., H. Murray, and J. P. Rushton. "Personality, Classroom Behavior, and Student Ratings of College Teaching Effectiveness: A Path Analysis," Journal of Educational Psychology, Vol. 77, 1985, pp. 394-407.

Feldman, K. A. "The Association Between Student Ratings of Specific Instructional Dimensions and Student Achievement: Refining and Extending the Synthesis of Data From Multisection Validity Studies," Research in Higher Education, Vol. 30, 1989, pp. 583-564.

Feldman, K. A. "Instructional Effectiveness of College Teachers as Judged by Teachers Themselves, Current and Former Students, Colleagues, and Administrators, and External (Neutral) Observers," Research in Higher Education, Vol. 30, 1989, pp. 137-194.

Greenwald, A. G. "Applying Social Psychology to Reveal a Major (But Correctable) Flaw in Student Evaluations of Teaching," Paper presented at a the annual meeting of the a American Psychological Association, New York City,

1995.

Marsh, H. "Students' Evaluation of University Teaching: Dimensionality, Reliability, 3 Validity, Potential Biases and a Utility," Journal of Educational Psychology, Vol. 76, a 1984, pp. 707-754.

Marsh, H., and M. Dunkin. a "Students' Evaluations of College Teaching: A Multidimensional Perspective," Higher a Education: Handbook of Theory and Research, Vol. 8,a 1992, pp. 143-233.

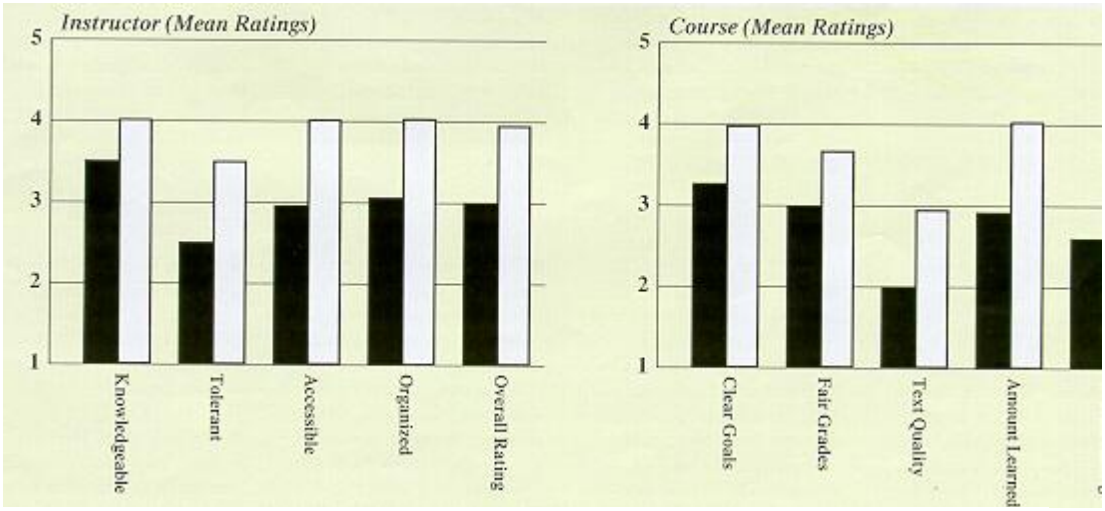
Woolfolk, A. E. and D. Brooks. "Nonverbal Communication in Teaching," in E. Gordon, ed., Review of Research in Education (Vol. 10, pp. 103-150), Washington, DC: AERA, 1983.

TABLE I. STUDENT RATINGS OF INSTRUCTOR AND COURSE FOR FALL AND SPRING SEMESTERS

	Semester			
	Fall (N = 299)		Spring (N = 243)	
Ratings of Instructor (1-5 Scale)	Mean	SD	Mean	SD
Knowledgeable	3.61	0.84	4.05	0.95
Tolerant	2.55	0.83	3.48	0.88
Enthusiastic	2.14	0.94	4.21	0.83
Accessible	2.99	0.90	4.06	0.82
Organized	3.18	0.90	4.09	0.88
Instructor mean	3.08	0.60	3.92	0.69
Rating of Course (1-5 Scale)	Mean	SD	Mean	SD
How much learned	2.93	0.98	4.05	0.91
Clear goals	3.22	0.92	4.00	0.96
Fair grades	3.03	0.89	3.72	0.98
Text quality	2.06	0.72	2.98	0.86
Overall course rating	2.50	0.91	3.91	0.93
Recommend course[*]	2.36	0.77	2.81	0.49
Course mean	2.70	0.62	3.65	0.76

Note: All between-condition comparisons of every variable are significant at the $p < .0001$ level.

[*] Scale for this variable was 1-3: 1 = No; 2 = Unsure; 3 = Yes.



S: CHART I. STUDENTS' EVALUATIONS AS A FUNCTION OF PROFESSOR'S VOICE PITCH (BLACK BARS = LOW VARIABILITY; WHITE BARS = HIGH VARIABILITY). (N = 472)

~~~~~~~~

BY WENDY M. WILLIAMS & STEPHEN J. CECT

Wendy M. Williams is an Associate Professor in the Department of Human Development at Cornell University, and the recipient of the 1996 Early Career Contribution Award in Educational Psychology from the American Psychological Association. She is the author of several hooks on educational practice and policy. Stephen J. Ceci is the Helen L. Carr Professor of Developmental Psychology at Cornell University. His past honors include a Senior Fullbright-Hayes fellowship and a Research Career Scientist Award. He is president of the division of General Psychology of A.P.A.
~~~~~~~~