Student Evaluations of Teachers

Students rate most highly instructors from whom they learn least.

Miriam Rodin and Burton Rodin

There are two ways to judge teaching through the medium of the students. The objective criterion of teacher effectiveness is based on what students have learned from the teacher. The subjective criterion is based on student evaluations of teacher effectiveness. The object of this study was to assess the validity of student evaluations by means of a comparison between the objective and subjective criteria of good teaching.

The subjective measure is rather easy to use. Students are asked to spend a few minutes evaluating the instructor's teaching performance. Good teaching is then defined as good scores on the student evaluation form (*1*). In 1960, 40 percent of institutions of higher learning were asking students to evaluate their teachers (*2*). There are no figures for 1970, but considering the growing focus on the teaching function of professors and the importance currently attached to student input, it seems safe to assume that there has been a substantial increase.

A criticism often leveled against the subjective criterion is that a student's evaluation of his instructor might depend on the grades he receives from that instructor. The standard reply is the citation of data (*3, 4*) which supposedly demonstrate that grades do not influence student ratings of instructors. Although extensively cited, these data do not in fact support this conclusion. Remmers asked 409 students to evaluate 11 different teachers in 17 different classes on the Purdue Rating Scale (*5*). The instructors then read off the names of the students in the top half of the class and asked those students to mark an X on their rating sheets. A biserial correlation between grades and ratings was then computed. Remmers found (*4*, p. 316) "correlations for individual traits of individual instructors varying from $-.860$ to $+.890$." He continued, ". . . the average of all correlations is $+.070$ at most. The conclusion seems inescapable, therefore, that for the average instructor and the average student there is practically no relationship between the student's grades and his judgment of the instructor. . . ." It is the last sentence which is always cited. However, taking the average correlation over traits and instructors is somewhat like characterizing the motion of a swinging pendulum as zero because the two directions cancel each other out. Contrary to the claim usually made of them, Remmers' data seem in fact to indicate that there is some relationship between grades and evaluations, although both the direction and the extent of this relationship vary from one instructor to the next. Thirty-six of the correlations, including both positive and negative ones, differed significantly from zero. Remmers *et al.* (*6*) later offered an interesting interpretation of these data. They suggested that some teachers direct their efforts to the poorer students and some to the better students. For the former a negative correlation between grades and ratings would be expected and for the latter, a positive one. The original table of data (*4*) offers support for this suggestion. The Purdue Rating Scale used had 10 subitems. If, contrary to the suggestion, the correlations between grades and subitem ratings vary randomly about a true correlation of zero, then for any instructor about half should be positive and half negative; 8 out of 10 in a consistent direction is an extremely unlikely outcome ($P < .05$). The data show that for instructors in 12 out of the 17 classes, at least 8 of the 10 subitem correlations were indeed in the same direction. Presumably, instruction in the remaining classes was pitched somewhere in the middle.

Two direct attempts to measure the correlation between objective and subjective criteria were made by Remmers *et al.* (*6*) and Elliot (*7*). Both used essentially the same experimental design. Students in a large introductory chemistry class were self-assigned to a number of different recitation and laboratory sections taught by graduate student teaching assistants. Using a multiple regression equation based on scores in various freshman placement tests, the experimenters predicted the average grade of each section. An instructor was defined as a good or a poor teacher according to whether the average grade for his section was above or below that predicted by the regression equation. This procedure corrects to some extent for differences in initial ability between sections. All instructors were rated by the students on a form of the Purdue Rating Scale specifically designed for use with Chemistry-1 instructors. Student grades were based on weighted averages from four 1-hour examinations, a number of short tests, and scores on laboratory and lecture notebooks. Remmers *et al.* (*6*) concluded that "there is warrant for ascribing validity to student ratings . . . as measured by what students actually learn of the content of the course." However, the data themselves seem inconclusive on this point. The instructors were rated on the Purdue Rating Scale for both laboratory and recitation sections; in this form of the rating scale, each section rating consisted of 12 subitems so that a total of 24 subitems was obtained. Five of these subitems significantly differentiated between good and poor instructors. Elliot (*7*), however, in a virtually identical study, was able to replicate significance for only one of these subitems ("Rating as compared to other instructors at Purdue University"). Furthermore, after having selected out the eight subitems found "differentiating most adequately" between good and poor instructors, Remmers *et al.* were only able to obtain a correlation of $+.266$. Elliot (*7*, p. 33) also concluded that "there is probably, in general, a positive relationship between the ratings given an instructor by his students and their achievement. . . ." He found the correlation between teacher ratings and student achievement to be $+.239$.

The results of the two experiments cited above do not justify the conclu-

Dr. Miriam Rodin is associate professor in the department of psychology, California State University, San Diego 92115. Dr. Burton Rodin is professor in the department of mathematics, University of California, San Diego, La Jola 92037.

sion that there is a positive relationship between the objective and subjective criteria: The observed correlations of + .266 and + .239 are not significantly different from zero. The crux of the problem of comparing the two criteria lies in obtaining an accurate measure of how much a student has learned. In our study, the objective criterion has been more carefully defined and controlled than heretofore.

Method

The instructors were teaching assistants in a large (293 students) undergraduate calculus course. All of the students met 3 days a week for a lecture by the professor in charge of the course. They met with individual teaching assistants in small recitation sections on the remaining 2 days. One recitation hour was devoted to answering questions about the lectures and homework. The other was devoted exclusively to administering test problems and going over preceding ones.

The course content was defined by 40 paradigm problems. Comprehension of a paradigm problem was tested with a specific problem. If the student missed that problem, he was allowed to retake variants of it (up to six times) until he passed. All of the students received a uniform sequence of variants constructed by the professor. The teaching assistants were not permitted to see the problems before the hour in which they were administered. The grading was done by the teaching assistants, but was completely objective: If any portion of the problem was done incorrectly, or if there was any error, no matter how trivial, the entire problem was scored as a miss. The final grade was completely determined by the number of problems passed. The number of attempts necessary to pass a problem had no effect on the grade.

The above procedure yields a fair and careful measure of what has been learned. In the first place, the test problems exhausted the content domain of the course. Most traditional examinations only sample the course content, and different samples are very likely to yield different estimates of how much was learned. Second, because of the opportunity for multiple administrations the measure of how much was learned was highly reliable. Scores on one-shot testing procedures are subject to a great deal of variability connected with factors such as personal problems, fatigue, gloomy compared to sunny rooms, and the like. The secrecy of the test problems was another important factor. If instructors have prior information about the contents of tests, they may, intentionally or unintentionally, pass it on to the students. It is always good practice to eliminate such potential bias from testing procedures. In this case it is particularly important since otherwise an obtained positive correlation may merely reflect the tendency of popularity-minded teaching assistants to hint at or "teach to" the test items.

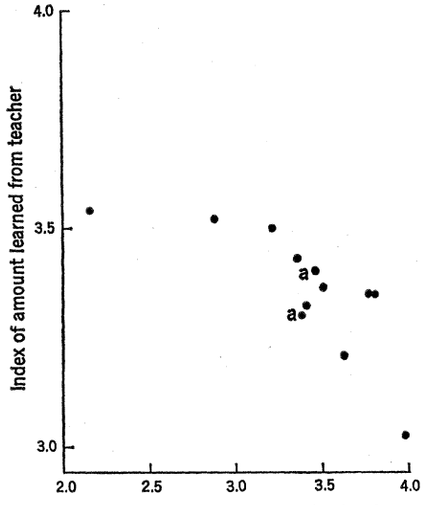

Fig. 1. Relationship between objective and subjective criteria of good teaching ($r = -.75$). The points labeled *a* are for two sections taught by the same instructor.

Criterion Measures

There were 12 sections, two of which were taught by the same instructor. Three measures were obtained for each of the sections: initial ability in calculus, amount learned by the students (objective measure of teacher effectiveness), and student evaluation of the instructor (subjective measure of teacher effectiveness). Since these data were collected in the third quarter of the course, a measure of the students' initial ability was available from their previous performance. Each section was assigned an initial ability score based on the mean grade obtained by the students in that section in the preceding quarter. (The students may or may not have had the same section instructor in earlier quarters.) The amount learned in each section in the current quarter was defined by the mean grade obtained in that section. The instructor for each section received a student evaluation score based on the mean rating given him by students in that section. The student evaluations of the instructors were made and collected during the large lecture section at the end of the quarter. Anonymous ratings were requested in the interest of obtaining honest responses. A number of subquestions were asked on the rating sheet. The question used in the analysis was "What grade would you assign to his total teaching performance?" (This is the question most similar to the one item which previous investigators found significant.) Numbers were assigned to letter grades in the usual way, with A equal to 4, B to 3, and so on, and with an adjustment of 0.5 made for borderline ratings.

Since students were allowed to choose recitation sections compatible with their preferences and schedules, the sections might be expected to vary in initial ability. Initial ability could affect the amount learned by the students. Therefore, in preference to a simple correlation between the objective and subjective measures of teaching effectiveness, a partial correlation between the two measures was obtained. This statistic "partials out" the effect due to initial ability, or, in effect, describes the relation between amount learned from the instructor and student rating of the instructor, with initial ability held constant.

Results and Conclusions

The partial correlation between the objective and subjective measures of teaching ability, with initial ability held constant, was equal to $-.746$. The probability is .95 that the true value of the correlation in the population, denoted by $\rho_{12.3}$, is covered by the interval $-.27 < \rho_{12.3} < -.93$. The ordinary correlation between the two measures was $r = -.754$.

Figure 1 is a scatter diagram for the two measures. Such a diagram does not correct for the differences in initial ability of the classes; nevertheless, it is useful because the ordinary and partial correlation coefficients happened to be nearly identical. The points labeled *a* represent the outcome for the two sections taught by the same instructor; their closeness is informal evidence for the reliability of the data. The instructors with the three lowest subjective scores received the three highest objective scores. The instructor with the highest subjective rating was lowest on the objective measure.

The apparent discrepancy between the

correlation reported here ($r = -.75$) and those reported by Remmers *et al.* [(*6*), $r = +.266$] and Elliot [(*7*), $r = +.239$] may be due to the procedure by which they obtained their objective measure. Although the examinations were objectively scored in their studies, the instructors had prior knowledge of the test questions. In addition, the grading of lecture and laboratory notebooks by individual instructors introduced a subjective and nonuniform element into the objective measure of the amount learned. In any case, these authors did not obtain a significant positive correlation between the two variables. The confidence intervals (as roughly estimated from their data) about their correlations would include negative values. In fact, although the result reported here contradicts the conclusions commonly drawn from Remmers *et al.* and from Elliot, it is not necessarily inconsistent with the data they obtained.

The explanation for the negative correlation between the amount learned from an instructor and the students' evaluation of his teaching performance is not obvious. Perhaps students do not wish so much to maximize the amount learned as to reach an equitable compromise between the effort involved in learning and the perceived importance of what is being learned. Or, in short, perhaps students resent instructors who force them to work too hard and to learn more than they wish. It may be that as students learn more, they become better able to detect the weaknesses of their instructors. Many other hypotheses could be advanced, but it seems fruitless to speculate without further evidence. Similarly, information about the extent to which the present results may be generalized to different types of courses must await future experimentation.

A correlation in the vicinity of .7 accounts for about one-half of the variance in student evaluation of their teachers. What accounts for the residual variance? There is evidence that student evaluations, to a large extent, tend to reflect the personal and social qualities of an instructor, "who he is" rather than "what he does" (*8*).

How should good teaching be measured? The major defense for defining good teaching in terms of good scores on the student evaluation forms is based on an analogy between the student and the consumer—the student, as the primary consumer of the teaching product, is in the best position to evaluate its worth. However, the present data indicate that students are less than perfect judges of teaching effectiveness if the latter is measured by how much they have learned. If how much students learn is considered to be a major component of good teaching, it must be concluded that good teaching is not validly measured by student evaluations in their current form.

References and Notes

1. V. Voeks, *J. Higher Educ.* **33**, 212 (1962); J. B. Bressler, *Science* **160**, 164 (19(→ J. R. Hayes, *ibid.* **172**, 227 (1971).
2. J. E. Stechlein, in *Encyclopedia of Educational Research*, C. Harris, Ed. (Macmillan, New York, 1960), pp. 286–287.
3. H. H. Remmers, *Sch. Soc.* **28**, 759 (1928).
4. ———, *J. Educ. Res.* **21**, 314 (1930).
5. The Purdue Rating Scale is a teacher evaluation form constructed by the education department of Purdue University. Students are asked to rate instructors according to various subitem traits, such as clarity of presentation, fairness, and so forth. Complete lists are given in (*3*, *4*, *6*, *7*).
6. H. H. Remmers, F. D. Martin, D. N. Elliot, *Purdue Univ. Stud. Higher Educ.* **66**, 17 (1949).
7. D. N. Elliot, *ibid.* **70**, 5 (1950).
8. R. H. Knapp, in *The American College*, N. Sanford, Ed. (Wiley, New York, 1962), pp. 290–311; K. Yamamoto and H. F. Dizney, *J. Educ. Psych.* **57**, 146 (1966).

9. Thanks to Dr. J. Anthony Deutsch for stimulating our interest in the topic and to Drs. A. Hillix and R. Penn for a critical reading of the manuscript. Supported by NSF grant GP-199872.

Public Interest Science

The governmental and public advisory activities of scientists have great political impact.

Frank von Hippel and Joel Primack

Although scientists as technical experts make important contributions to the federal policy-making process for technology, that process remains basically political. At present, the primary recipient of technical advice on matters of public policy is the executive branch of the federal government. To the extent that this arrangement results in an informed executive branch dealing with a relatively uninformed Congress and public, a corresponding shift in power occurs. Indeed, it is not unheard of for the executive branch to abuse its near monopoly of politically relevant technical information and expertise. We cite below several case studies exemplifying the sorts of abuses that occur: politicization of advisory committees; suppression and misrepresentation of information, and analyses.

This leads us to the question of whether individual scientists can contribute significantly to a restoration of a balance of power between the public, Congress, and the executive branch of the government. We find, again on the basis of case studies, that a few scientists can be surprisingly effective in influencing federal policies for technology if they are sufficiently persistent and skillful and if various other circumstances are favorable. These success stories and the present high level of concern about the adverse side effects of technology among both scientists and the public suggest that the time is propitious for a much more serious commitment within the scientific community to "public interest science."

This article is divided into two main sections. The first deals with devices by which the executive branch exploits its scientific advisers for political advantage while concealing much of the information they have provided; the second discusses ways in which scientists

Dr. von Hippel is an associate physicist at Argonne National Laboratory, Argonne, Illinois. Dr. Primack is a junior fellow of the Society of Fellows and a member of the physics department at Harvard University, Cambridge, Massachusetts.

This article is adapted from an invited talk that was given by Dr. von Hippel at the annual meeting of the American Physical Society, January 1972.