1995

Student evaluations of teaching effectiveness: the interpretation of observational data and the principle of faute de mieux

B. Burt Gerstman, *San Jose State University*

Gerstman, B. B. (1995). Student evaluations of teaching effectiveness: The interpretation of observational data and the principle of *faute de mieux*. *Journal on Excellence in College Teaching*, *6*(3), 115-124.

Student Evaluations of Teaching Effectiveness: The Interpretation of Observational Data and the Principle of *Faute de Mieux*

B. Burt Gerstman
San José State University

Student opinion surveys are important but widely misunderstood tools for evaluating teaching effectiveness. In this brief review, an analogy is drawn between the use and interpretation of observational data for public health and biomedical research and the use of student opinion data in evaluating teaching effectiveness. Sources of systematic error in the form of selection bias, information bias, and confounding are defined and illustrated. Original data concerning intermittent "quid pro quo" confounding (i.e., the effect of expected grades on student evaluations of teaching) are presented. Finally, the principle of faute de mieux *("lack of anything better") and the interpretation of less-than-pristine data are considered.*

Introduction

Nearly everyone in higher education has an opinion about the value of student evaluations of teaching effectiveness. Without question, these evaluations are among the most important sources of information considered by university retention, tenure, and promotion committees and university administrators alike. In my opinion, there has been some degree of misinterpretation and misuse of these data, which has resulted in controversies similar to those that surrounded my field, epidemiology and biostatistics, not too long ago.

A turning point for the disciplines of epidemiology and biostatistics occurred in 1964, when the Surgeon General of the United States convened a panel of scientists to advise him on the effects of cigarette smoking

on health. The panel adopted a set of criteria that formalized and legitimized the use of nonexperimental data for causal inference (U.S. Department of Health, Education & Welfare, 1964). These criteria have helped clarify epidemiologic debates ever since.

Epidemiology has matured as a science over the last 30 years, through the establishment of methods to increase the accuracy of observational data collection, analysis, and interpretation. One critical aspect of epidemiologic methods is the control and mitigation of bias. In this article, I will discuss the basic concepts of bias and demonstrate how they may affect student evaluation data. I will also discuss the principle of *faute de mieux*, which translates roughly as "lack of anything better," in relation to student evaluations. That is, as imperfect as student evaluations might be, they still provide important insights into teaching performance. It is true that student evaluation data are compromised by the lack of potentially relevant information about the students, the uncontrolled circumstances under which student evaluation data are collected, and the lack of objectivity associated with student opinion. However, no other observers—whether peers; retention, tenure, and promotion committee members; or university administrators—have greater opportunity to observe and assess a professor's performance than do students. Therefore, student evaluations of teaching effectiveness probably will continue to be a valuable source of information in assessing teaching performance at the university level.

Before proceeding, I should note that significant discoveries regarding health have been made on the basis of relatively crude, uncontrolled data (e.g., the adverse effects of smoking on health). However, the collection and interpretation of these data have been fundamentally different from the collection and interpretation of "clean" experimental data. The challenge is to recognize the limitations inherent in "dirty" data and to apply only those means of analysis, interpretation, and inference that are appropriate. To understand the "dirtiness" of data, fundamentals of study design must be considered first.

Differences Between Experimental and Observational Data

There are two different types of data: experimental and observational. The distinction between the two is clear: In experiments, investigators control the allocation of the factors ("treatments" or "independent variables") they wish to study; in observational studies, no such discretion is allowed—selection of the study factors is in the hands of the observational study participants themselves. Control over study factor allocation

by investigators permits randomization of treatments. Randomization, in turn, tends to result in group comparability, making the comparisons that follow fairly straightforward. In the absence of randomization, comparisons may be confounded by the underlying noncomparability of groups. Therefore, experimental study designs are almost always preferable to observational designs when feasibility and ethical considerations allow a choice between the two.

Less distinct differences between experimental and observational studies are also apparent. For example, experimental studies are often characterized by homogeneity of study participants, whereas those participating in observational studies may be quite diverse. The clearest example of experimental study participant homogeneity is in laboratory and agricultural experiments where treatments are randomized among genetically identical laboratory animals or crops. Even in experimental studies involving humans, participant homogeneity can be effected through strict admissibility criteria for participant selection.

Finally, experimental studies are often carried out in relatively controlled environments. In contrast, observational studies are pursued in natural, population-based settings where environmental and other extraneous conditions are heterogeneous and uncontrolled. This can further confound comparisons in that observed differences may be due to environmental factors other than the independent variable under investigation. For all of the aforementioned reasons, comparisons based on observational studies must be controlled, statistically adjusted, or otherwise compensated for before being interpreted or scrutinized.

Acceptability of Nonexperimental Data

R. A. Fisher (whom many consider the "father of modern statistics") and other skeptics scorned inferences based on nonexperimental data. In addition, Fisher believed that solutions derived without a full understanding of the sort of reasoning behind experiments were unjustifiable. He suggested that statistics be entrusted only "to those with sufficient prolonged experience of practical research, and of responsibility for drawing conclusions from actual data upon which practical action is to be taken" (Box, 1978, p. 435).

Until relatively recently, the accepted norm in the conservative realm of science was not too different from Fisher's curmudgeonly views. The Report of the Advisory Committee to the Surgeon General of the Public Health Service (U.S. Department of Health, Education & Welfare, 1964) went some way toward reversing this excessively orthodox view. Importantly, the Surgeon General's report did agree with Fisher's basic

premise that inference is a matter of judgment which goes beyond numerical manipulations. Still, the Surgeon General's report went beyond this generality and actually documented nonstatistical criteria by which to judge causality.[1] The more pertinent criteria were the following: (a) Cause-and-effect conclusions based on a single study are rarely justifiable; rather, consistency between studies using diverse methods in different populations, but providing largely similar conclusions, is necessary for reliable conclusions; (b) large differences carry more weight than do small differences; and (c) statistical associations must be plausible and coherent in the face of other known facts relevant to the topic. (Other criteria outlined by the Surgeon General's report were more specific to biological phenomena.)

By applying this thinking to the interpretation of student evaluations, the inferences based on these data are strengthened.

Sources of Inaccuracy

In assessing data of any type, two different sources of error must be kept in mind: imprecision and bias. Imprecision is synonymous with "random error" and often is associated with small sample size and the resultant sampling variation. Imprecision causes the random noise that must be subdued to determine whether apparent differences are real or random. This noise is easily quantified in terms of standard error estimates and confidence interval lengths. Moreover, it can be handled intelligently by means of statistical significance testing.

Bias, the other major form of error, is any condition that tends to create systematic deviations from the truth (Sacket, 1979). Bias repeatedly leads to the wrong conclusions and is to some degree independent of random sources of error. Bias, more than imprecision, is of central concern when dealing with observational data.

According to current epidemiologic theory, there are three principal forms of bias: information, selection, and confounding. Although these three forms of bias are distinct by definition, they tend to overlap in practice. Nonetheless, it is worthwhile to consider them separately to help define and clarify their potential adverse effects.

Selection Bias

Selection bias occurs when the study participants do not fully represent the population that they supposedly represent. In student evaluations, selection bias happens if some professors have the option of surveying only their more agreeable classes, whereas other professors do

not. This practice tends to extend the "normal" range. Comparisons (formal or otherwise) that follow therefore may be biased in favor of self-selected classes.

Selection bias may be prevented simply by taking the class selection process out of professors' hands. At the very least, comparisons could be based on a random sample of classes taught at the university.

Information Bias

Information bias occurs when one or more variables are misclassified or inaccurately measured. For the purpose of this discussion, student ratings of teaching effectiveness will be the primary outcome, or dependent variable, of the study.

The quintessential question about effectiveness ratings is how well they reflect the complex interaction of teaching and learning. Although a thorough treatment of this question is beyond the scope of this general overview, this issue points to the importance of objectively defining study outcomes and endpoints before referring back to results.

In population-based studies of disease occurrence, criteria that define study outcomes are called "case definitions." Accurate case definitions are essential to study reliability and acceptability. In my view, the case definition of student evaluations of teaching effectiveness can be viewed in one of two ways: as a surrogate endpoint for what classically have been the goals of effective teaching—stimulating intellectual curiosity, developing thought processes and critical thinking skills, and preparing students to contribute to society (Cruse, 1987)—or as an outcome in and of itself. Unfortunately, the distinction between these two views is often disregarded.

If student evaluations are viewed as true reflections of classical learning objectives, they are, at best, surrogate or substitute measures. As with all surrogate measures, there is ample opportunity for information bias. If student evaluations are viewed more literally, such as a measure of rapport with and ability to please students, information bias is less of an issue. Machina (1987) argued that reaching students may be considered a prerequisite for effective teaching and therefore is an acceptable outcome.

Confounding

Confounding is a result's distortion by extraneous variables. For an extraneous variable ("potential confounder") to confound, it must be associated with both the independent and dependent variables under consideration (Rothman, 1986).

Extraneous variables that may confound student evaluations include, but are not limited to, the reason for taking the class (Brandenburg, Slinde, & Bautista, 1977; Feldman, 1978); discipline (Centra & Creech, 1976); class size (Centra & Creech, 1976; Feldman, 1978; Marsh, Overall, & Kesler, 1979); and expected grade (Centra, 1979; Feldman, 1983). Over the past few years, I have become increasingly interested in this last form of confounding that may result from different grading standards. I shall refer to this hypothesized effect as "quid pro quo" confounding.

Quid pro quo confounding may occur when relatively lenient grading standards are associated with higher-than-average student evaluations and when relatively stringent grading standards are associated with lower-than-average student evaluations. To quantify the hypothesized quid pro quo effect, I conducted a simple study in which two of my classes were surveyed (it is fair to say that class 1 was relatively less challenging and had less stringent grading standards than class 2). Both classes were administered a brief survey in which students were asked to rate my overall teaching effectiveness on a scale of 1 to 5 (1 = excellent, 2 = good, 3 = fair, 4 = poor, and 5 = very poor). Students were also asked to indicate the grade they expected to receive on the basis of the exams and assignments to date. Students did not identify themselves on the survey, so analysis of the data was blind.

Interestingly, in class 1 there was no relationship between expected grade and teaching effectiveness ratings (Figure 1, $p = .38$). In class 2, on the other hand, there was a strong linear relationship (Figure 2, $p = .006$), such that the higher the expected grade, the higher the rating of teaching effectiveness. This inconsistency of effect indicates a statistical interaction. It appears that higher grading standards and more rigorous materials are associated with quid pro quo confounding, whereas a more laid-back approach is not. I believe this result has far-reaching implications for higher education.

On a related note, I would like to point out what I suspect is a fallacy in the educational literature concerning the potential effects of confounding. Marsh (1984) suggested that student ratings are "relatively unaffected by a variety of variables hypothesized as potential biases" (p. 707). The statistics behind this and similar statements in the literature are coefficients of determination or other standardized regression or correlation coefficients. For example, Seldin (1993) stated that "relationships between extraneous variables and student ratings . . . account for just 12 to 14 per cent of the variance between positive and poor ratings" (p. A40). The statistic in this case ("12 to 14 per cent of the variance") appears to be a coefficient of determination (abbreviated in the statistical literature as r^2). I believe that this statistic is inappropriate, given the nature of stu-

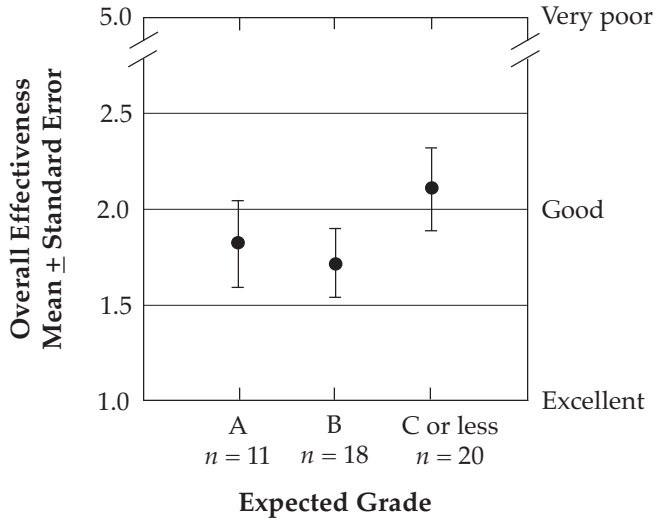

Figure 1. Mean ± standard error of student ratings of teaching effectiveness by self-estimated student grade, class 1 (less challenging curriculum and less stringent grading standards).

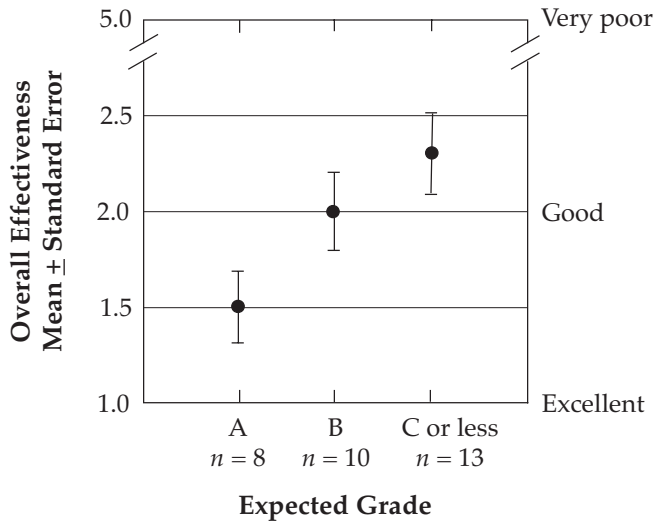

Figure 2. Mean ± standard error of student ratings of teaching effectiveness by self-estimated student grade, class 2 (more challenging curriculum and more stringent grading standards).

dent evaluation data. Although coefficients of determination provide an intuitive measure of an independent variable's effect in reducing variance in the dependent variable, they are generally inappropriate for causal inference based on observational data (Rothman, 1986, p. 303). This is because coefficients of determination depend, in part, on the overall frequency and range of covariates in question. Greenland, Schlesselman, and Criqui (1986) therefore suggested that coefficients of determination and other standardized correlation parameters be avoided as an analytic tool in observational studies of cause and effect. Another problem with coefficients of determination is that they represent average correlations and are therefore insensitive to interactions of the type illustrated in the aforementioned simple study. Further consideration of the inadequacy of coefficients of determination and similar coefficients for causal inference based on observational data is beyond the scope of this commentary. Interested readers should see Neter, Wasserman, and Kutner (1985, especially p. 99) and Greenland et al. (1986) for a more technical discussion of this matter.

Faute de Mieux

Given the nonrigorous nature and the vagaries of student evaluation data, their value might be questioned. A recent debate concerning the use of nonrigorous biomedical data for epidemiologic research may shed some light on this subject. Feinstein (1989) referred to the logic behind the use of nonrigorous biomedical data as the *faute de mieux*, or "lack of anything better," reason. He eloquently expressed the essence of this debate as follows: "Caught in the necessity for making decisions based on evidence and the pragmatic difficulty of getting high quality evidence, . . . investigators and policy makers usually conclude that imperfect data are better than none" (p. 930).

The similarities between the use of nonrigorous biomedical data and the use of student evaluation data are obvious yet striking. Academic administrators, when looking for evidence of teaching effectiveness as they consider the promotion, tenure, or reappointment of faculty, often rely on the quantification inherent in student evaluation data.[2] Unfortunately, administrators may mistake the numerical nature of the data for objectivity (it is seductive to confuse the two). Without an appreciation of the potential sources of error inherent in student evaluation data, the inferences that follow may be invalid and the data potentially abused. The potential for selection bias, information bias, and confounding must be considered, and adjustments must be made, if possible. (As far as I know, very little has been done in advancing statistical adjustment for

student evaluation data.) Moreover, student evaluations must be viewed in light of other sources of information, and small differences should not be overinterpreted. Most important, statistical data must be plausible and coherent in the face of other relevant facts. There is a saying in the epidemiologic community attributed to Michael Gregg, "We are always dealing with dirty data. The trick is to do it with a clean mind" (Bernier & Mason, 1991, p. 236). Consumers of student evaluation data would be well served to adopt a similar philosophy.

Footnotes

[1]It should be noted that the Surgeon General's Advisory Committee based their criteria on the earlier work of Sir Bradford Hill. See Hill (1965) for a description of the inferential criteria he originally proposed.

[2]As part of a recent retention, tenure, and promotion (RTP) grievance hearing at my university, I was asked to blind-review the student ratings of a professor's teaching effectiveness and to render an interpretation. Apparently, the case revolved around a small number of student ratings that had fallen below the normative range. Few supportive data were offered. This points to the overreliance on student evaluations as the de facto "gold standard" of many RTP actions.

References

Bernier, R. H., & Mason, V. M. (Eds.). (1991). *Episource: A guide to resources in epidemiology*. Rosewell, GA: The Epidemiology Monitor.

Box, J. (1978). *R. A. Fisher: The life of a scientist*. New York: Wiley.

Brandenburg, D. C., Slinde, J. A., & Bautista, E. E. (1977). Student ratings of instruction: Validity and normative interpretations. *Journal of Research in Higher Education*, *7*, 67-78.

Centra, J. A. (1979). *Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement*. San Francisco: Jossey-Bass.

Centra, J. A., & Creech, F. R. (1976). *The relationship between students, teachers, and course characteristics and student ratings of teacher effectiveness* (Project Report No. 76-1). Princeton, NJ: Educational Testing Service.

Cruse, D. B. (1987). Student evaluations and the university professor: Caveat professor. *Higher Education*, *16*, 723-737.

Feinstein, A. R. (1989). Para-analysis, faute de mieux, and the perils of riding on a data barge. *Journal of Clinical Epidemiology*, *42*, 929-935.

Feldman, K. A. (1978). Course characteristics and college students' rating of their teachers: What we know and what we don't. *Research in Higher Education*, *9*, 199-242.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education*, *18*, 3-124.

Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology*, *123*, 203-208.

Hill, B. A. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*, 295-300.

Machina, K. (1987, May-June). Evaluating student evaluations. *Academe*, 19-22.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707-754.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal*, *16*, 57-70.

Neter, J., Wasserman, W., & Kutner, M. H. (1985). *Applied linear statistical models: Regression, analysis of variance, and experimental designs*. Homewood, IL: Richard D. Irwin.

Rothman, K. J. (1986). *Modern epidemiology*. Boston: Little, Brown.

Sacket, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, *32*, 51-63.

Seldin, P. (1993, July 21). The use and abuse of student ratings of professors. *The Chronicle of Higher Education*, p. A40.

U.S. Department of Health, Education & Welfare. (1964). *Smoking and health: Report of the Advisory Committee to the Surgeon General of the Public Health Service* (Public Health Service Publication No. 1103). Washington, DC: U.S. Government Printing Office.

B. Burt Gerstman *is Associate Professor of Health Science at San José State University. He has published extensively on the safety of oral contraceptives and other pharmaceutical drugs. He currently serves as chair of his university's Student Evaluation Review Board.*