Widener University Delaware Law School

From the SelectedWorks of Larry D Barnett

1996

Are Teaching Evaluation Questionnaires Valid? Assessing the Evidence

Larry D Barnett

Available at: https://works.bepress.com/larry_barnett/24/

J. COLLECTIVE NEGOTIATIONS, Vol. 25(4) 335-349, 1996

ARE TEACHING EVALUATION QUESTIONNAIRES VALID? ASSESSING THE EVIDENCE

LARRY D. BARNETT
Widener University, Wilmington, Delaware

ABSTRACT

Because the results of student-completed teaching-evaluation questionnaires often play a role in personnel decisions made by institutions of higher education, the article reviews the principal quantitative studies that social scientists have conducted on the validity of the questionnaires. The methodological flaws in and limitations of these studies are considered, and the potential side effects of the questionnaires are discussed. The article suggests that, because the questionnaires have not been shown to measure teaching quality accurately, use of the questionnaires in promotion, retention, and tenure decisions potentially violates the employment contract of faculty members.

Student-completed teaching evaluation questionnaires are a fact of life in institutions of higher education today, and they frequently play a major role in decisions on retention, promotion, and tenure. The questionnaires are ostensibly used to measure the quality of teaching, a use that is administratively justified by reference to the employment contract between faculty members and their institutions: the contracts of tenure-track and tenured faculty members require evaluation of the caliber of instruction provided by the faculty member. However, the questionnaires would seem to be supportable on contractual grounds only insofar as they *correctly* assess teaching quality [1]. While a faculty member who challenges a personnel decision has the burden of proving the inaccuracy of a performance appraisal by her/his institution, claims filed by current or former employees against their employers have increased rapidly since 1980 [2, pp. 1614, 1628 n.164]. It is consequently logical to expect that, if student-completed evaluation questionnaires cannot be shown to measure what they claim to measure (namely, teaching quality), college and university faculty members will start to contest

© 1996, Baywood Publishing Co., Inc.

negative decisions regarding retention, promotion, and tenure that are based in substantial part on data from the questionnaires. Alternatives to the questionnaires, after all, can be developed to gauge teaching effectiveness [3].

The validity of the questionnaires is of course a matter within the expertise of social scientists, not lawyers, and unfortunately the methodology of social science research is understood by few in the legal profession. Being a social scientist, I was therefore interested in two summaries of social science research on the questionnaires that were recently brought to my attention [4], and I decided to review the most prominent quantitative studies on which the summaries appeared to rely for their conclusion that the questionnaires are useful for the assessment of instructional merit.

Like issues in law, issues in social science are generally much more complex than they appear on the surface, and an examination of social science research must go beyond the findings that investigators report. In the pages that follow, accordingly, I will explore several technical issues in social science that are central to research on the questionnaires. My discussion necessarily involves abstract and complex concepts, but I have tried to make them understandable.

MEASURING INSTRUCTIONAL QUALITY

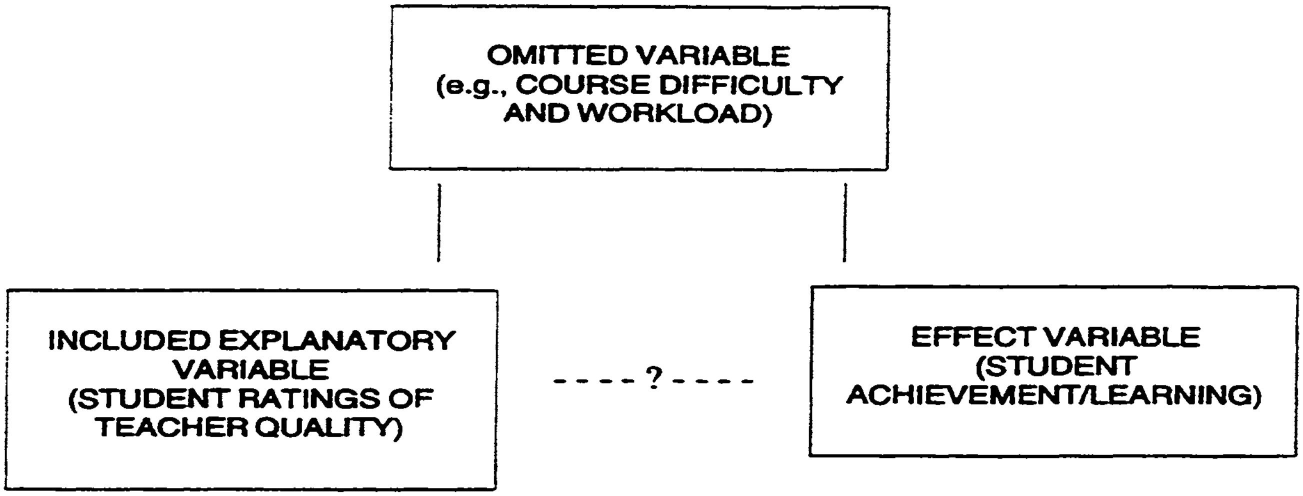
I begin with an obvious point: The objective of student-completed teaching-evaluation questionnaires is to measure the quality of teaching. This simple point, however, presents a researcher with an array of difficult problems and requires a number of assumptions. In ascertaining the degree to which teaching evaluation questionnaires correctly assess instructional quality, a criterion is needed to judge teaching quality. The main criterion, which is based on the assumption that the caliber of teaching is reflected in how much students learn, has been the extent to which students have mastered the material in their courses. Clearly, the assumption is not unreasonable, but we need to remember that the researcher is not directly capturing the quality of teaching. Rather, instructional quality is being estimated with an indicator (i.e., student learning), and the unavoidable result is measurement error. In addition, the extent to which students have mastered material is almost always ascertained by the performance of the students on tests administered in their courses, and the assumption that the tests flawlessly gauge what has been learned is hardly tenable. The second assumption, like the first, thus constitutes a source of measurement error in social science research on the validity of teaching evaluation questionnaires.

Currently, the assessment of instructional quality involves an unknown amount of error in measurement, but social scientists proceed on the premise that such error is minimal. This premise is troubling and raises the first serious question about research on teaching-evaluation questionnaires: if researchers are wrong in believing their work suffers from little measurement error, their findings are suspect. The danger has been ignored, however, and researchers have focused on whether and to what degree a relationship exists between the ratings students assign their teacher in a course and the score (or grade) students receive on their course examination(s). To the extent the students who favorably evaluate a teacher perform well on tests and the students who unfavorably evaluate a teacher perform poorly on tests, student-completed questionnaires are presumed to be accurate (valid) instruments for measuring instructional quality. The concern of social science research in this area, then, is with the strength of the association between the variable of student responses on teaching-evaluation questionnaires and the variable of student achievement on examinations.

STATISTICAL TESTING FOR RELATIONSHIPS BETWEEN VARIABLES

How does a social scientist determine whether variables are related and, if they are, whether the relationship is strong or weak? The most common approach is to analyze quantitative data with one of two statistical techniques, viz., correlation or regression. While the mathematical properties of regression have been more fully developed and regression is more often employed in social science research, correlation remains a powerful tool. Regardless of which of the two techniques is used, however, it is important to keep in mind that neither technique will supply an accurate estimate of a relationship if the investigators seriously err in measuring their variables, designing their research, or analyzing their data.

At this point I want to consider an error of a statistical nature that investigators may commit. The error arises from the logic of data analysis, and understanding it may be aided by the diagram below. In the diagram, our main concern is with the horizontal (dashed) line, and the question is whether the line represents a real relationship. In other words, we seek to determine whether a variable whose behavior we wish to explain (i.e., the "effect variable") is related to a variable (the "explanatory variable") that we believe accounts for the behavior of the effect variable and that we have therefore included in the statistical analysis [5]. In making this determination, we must consider the possibility that a correlation or regression coefficient indicating the presence of a relationship between the explanatory variable and effect variable is merely the product of a variable that has been omitted from the analysis. To the extent an omitted variable creates or inflates a coefficient for the relationship between the explanatory and effect variables, our data analysis will be faulty, and any conclusion based on it will be mistaken. We avoid an incorrect result by bringing the omitted variable into the analysis and statistically controlling (i.e., removing) its influence. With the previously omitted variable controlled, we can conclude that the explanatory variable and the effect variable are in fact associated, and hence that the former explains at least partly the behavior of the latter, if the correlation or regression coefficient for their relationship is not zero [6].


OMITTED VARIABLE (e.g., COURSE DIFFICULTY AND WORKLOAD)
INCLUDED EXPLANATORY VARIABLE (STUDENT RATINGS OF TEACHER QUALITY)
- - - - ? - - - -
EFFECT VARIABLE (STUDENT ACHIEVEMENT/LEARNING)

Not all variables that are omitted, however, will produce seriously biased results and wreck a study. Under what conditions must an omitted variable be included in a statistical analysis? The answer is that, if an omitted variable is related ***both*** to the explanatory variable ***and*** to the effect variable, it must be incorporated into the analysis so that its influence can be removed. This point cannot be overemphasized. The missing variable in such a situation can create a grave error not only with respect to the magnitude of a relationship reported by the researcher but also with respect to the direction of the relationship: even though a correlation or regression coefficient may indicate the presence of a relationship between the explanatory variable and the effect variable, no relationship may in fact exist, or the actual relationship may be opposite in sign (e.g., negative rather than positive). The following passage, which explains the problem in terms of regression, may help to understand the statistical mechanisms involved:

> What is happening when a relevant variable is excluded from the estimation of a regression model is that the [explanatory] variables left in the regression equation that are correlated with the excluded variable will pick up some of the impact of the excluded variable on [the effect variable]. The result is biased estimators of those variables left in the equation. The ***direction*** of the bias (positive or negative) will depend on both the direction of the effect of the excluded variable on the [effect] variable, and the direction of the relationship between the included and excluded variables. The ***magnitude*** of the bias depends directly on the relationship between the included and excluded variables: The more highly related the variables are, the greater will be the bias [7, p. 21].

I turn now to a second statistical matter—the interpretation of correlation coefficients. The numerical value of a correlation coefficient can be deceiving, even when it is "statistically significant" (i.e., even when it is unlikely to have occurred by chance if no relationship exists). A correlation coefficient measures the degree to which change in the amount of the explanatory variable is accompanied by change in the amount of the effect variable, but the most beneficial feature of a coefficient is not its numerical value, which has no inherent, practical meaning. Rather, the *square* of the numerical value is the most advantageous aspect of a correlation coefficient, for the square indicates the proportion of variation in the effect variable that can be statistically attributed to variation in the explanatory variable. The research summary by Professor Cashin reported a correlation coefficient of .44 between student ratings of an instructor "overall" and examination grades [8]. This coefficient means that 19.4 percent of the variation in student learning (as measured by course grades) is explained by variation in instructional quality (as measured by student ratings). If accurate, a correlation of this magnitude is "practically useful," as Professor Cashin said, though one must keep in mind that four-fifths of the variation in course grades remains unexplained and is attributable to other factors.

RESEARCH ON QUESTIONNAIRE VALIDITY: MULTISECTION STUDIES

But does this correlation coefficient accurately estimate the relationship between student evaluations of teaching and student achievement? The best research on the magnitude of the relationship is the "multisection validity study." When it is ideally designed, such a study possesses the following features: each course included in the study has numerous sections; students are randomly assigned to sections; the sections of a course have different instructors but a common textbook and the same examination(s); all examinations for a course are constructed by a person who does not teach any section; and subjective (essay) components of examinations are graded by the person who developed them. A review of multisection validity studies cites one work that, the author of the review asserts, eliminates at least in part "many of the criticisms of the multisection validity study" and "provide[s] strong support for the validity of students' evaluations of teaching effectiveness" [9, p. 721]. However, the cited work—which subjected the results of other multisection studies to a statistical analysis—did not control a number of critical variables that could have generated or enlarged the relationship between student ratings of teachers and student achievement [10]. Among the missing variables that might have explained the relationship was the rigor of the requirements of the instructor (such as checks for student preparation and amount of material assigned), a factor that may vary considerably across sections of a single course. If the variable was related both to student ratings of instructional quality and to student achievement, a control for the variable could have markedly weakened or entirely eliminated the relationship originally found between student ratings and student achievement [11].

Another variable that the work omitted was the students' level of interest in the subject matter of the course prior to exposure to the teacher they later evaluated. As will be suggested below, neither of these variables should have been excluded from the analysis and left uncontrolled.

While the work did not incorporate a number of potentially important variables into its data analysis, the work is the source of a set of correlation coefficients (including the coefficient of .44) that Cashin suggested are credible estimates of the relationship between student ratings of teachers and student achievement. A reader of the reproduced coefficients can easily be misled, however, because Cashin failed to make clear that the coefficients may have been seriously confounded by variables whose influence was not removed. The failure to clarify this point is surprising inasmuch as Cashin explicitly stated that a control is necessary for one of the variables omitted by the work, namely, the interest students initially exhibited in the subject [12, p. 5].

Let me turn to the other variable I mentioned that was not included in the work, namely, course difficulty and workload. Cashin contended that the variable is not in need of control, but he reached the conclusion evidently because he found the strength of the relationship between course difficulty/workload and student ratings of instructional quality to be just "modest" and because the direction of the relationship is such that "students give *higher* ratings in difficult courses where they have to work hard" [4, p. 6; emphasis in original]. Unfortunately, Cashin overlooked the relationship between difficulty/workload and student achievement (the effect variable). If increases in difficulty/workload raise not only student ratings but also student achievement, a control for difficulty/workload can appreciably reduce the coefficient for the association of student ratings and student achievement. Although there have been contrary findings, studies often report that students learn more in courses they perceive to be difficult, and the correlation coefficients frequently are at least "modest" in size [13, pp. 600-601]. Similar data appear in a monograph coauthored by Cashin [14]. As a consequence, course difficulty/workload cannot be disregarded in determining whether and how student ratings are associated with student achievement.

In another multisection study, student judgments of "course difficulty and workload" were measured, but the variable was not controlled in calculating the correlation coefficient for the relationship between student ratings of teacher effectiveness and student scores on final examinations [15]. The investigator may not have controlled the variable because, in six of the seven courses studied, its correlation with final examination scores was not statistically significant and hence could be attributed to chance. Given the research design of the study, however, statistical significance was a dubious criterion for decisions regarding the analysis of data. A central aspect of the research was that questionnaires were completed anonymously and the responses of an individual student could not be linked to the final examination score of the student. Because of this, correlation coefficients could not be calculated with data on each student; instead, coefficients were computed using the mean (i.e., arithmetic average) of each section for each of the three variables (viz., student judgments of course difficulty/workload, student ratings of instructional quality, and student examination scores). The number of observations was thus the number of sections, not the number of students on whom information was obtained, and the number of sections was small: five of the seven courses studied had fewer than nine sections, and no course had more than twenty-two sections. As I explain below, these twin features of the research are critical, for they compromise the conclusions that can be drawn from the study.

In assessing the validity of student ratings of instructors, data on individual students are clearly preferable to section averages, but unfortunately almost all research on the question employs section averages [16]. However, there are at least three reasons for preferring a statistical analysis based on individual-level data. One is that the ratings are concerned with individuals; indeed, the purpose of the ratings is to ascertain and improve the extent of learning by *individuals*. Another reason is that, as sociologists have long been aware, correlation and regression coefficients calculated with group-level measures (e.g., section means) do not necessarily correspond to coefficients calculated with individual-level measures [17]. A third reason is that, when the unit of observation is the individual rather than the group, the number of observations will be larger. This point is important to the research we are considering because the level of statistical significance achieved by a correlation or regression coefficient is sensitive to the number of observations on which the coefficient is based: to account for the role of chance, a coefficient that is calculated on a small number of observations must possess a far larger numerical value to reach a specified level of statistical significance than a coefficient that is calculated on a substantial number of observations.

Let me elaborate on the last point. The impact of number of observations on the ability of a coefficient to attain a given level of statistical significance strongly suggests that, where sections are the unit of observation in questionnaire validation studies and the number of observations (sections) is small, statistical significance is a tool of doubtful utility. The argument against the use of statistical significance in these studies is reinforced by the manner in which their observations are selected: probability sampling (i.e., sampling that provides each member of the eligible population with a known, nonzero chance of selection) seems not to be employed even though such sampling is at the foundation of the concept of statistical significance. Since mistaken findings become more likely to the extent that elements of research design and data analysis are unsound, significance levels should not influence the analysis of data obtained from group-level measurement and nonprobability sampling.

In the study we are presently considering, therefore, the judgment made by students of course difficulty/workload was a potentially important variable to control, regardless of the statistical significance of the coefficient for its correlation with scores on the final examination. This coefficient, I note, had a numerical value of .10 or higher in each of the seven courses studied and a numerical value of .30 or higher in three of the courses. If the variable of course difficulty/workload was related also to student ratings of teacher quality with a coefficient of the same sign—a result that was not reported by the investigator—it could well have inflated the coefficient for the relationship between the ratings and examination scores. Had the variable been controlled, therefore, the coefficient for the latter correlation might have been appreciably lower in many if not all of the seven courses.

What if, after controlling all relevant variables, the evaluation of teachers by students is found not to be statistically associated with the grades the students receive in a course? Some studies report no association, and the research summary by Professor Felder cited the studies to justify the conclusion that student evaluations of teaching constitute valid measures of instructional quality because the evaluations are not "just popularity contests" [4]. Such reasoning, however, misapprehends the logic of research on the validity of student ratings. In this research, student grades are treated as an indicator of the degree to which students have learned the material covered in their courses, and the amount of learning is in turn assumed to manifest the quality of teaching to which students are exposed. However, as the relationship weakens between the ratings teachers receive from their students and the grades students receive from their teachers, the questionnaires completed by students are deemed not to be an accurate (i.e., valid) measure of instructional quality.

If there is no statistical link between student ratings and student grades, then responses to evaluation questionnaires are not portraying the extent of student learning or, by implication, the effectiveness of teaching. Studies that find student evaluations and student grades are not associated, in other words, are directly counter to the proposition that students accurately describe the caliber of instruction. Notably, Felder, in defending the proposition with this finding, is explicitly contradicted by one of the studies he cited for support. Specifically, the investigators who carried out the study concluded that:

> What is being measured by student evaluations of teaching effectiveness remains an open question and a disturbing one. Our findings lead us to believe that students may evaluate instructors on the basis of somewhat subjective feelings that are not related in any direct way either to the grades they receive or to *how much they learn from the instructor* [18, p. 862; emphasis supplied].

Moreover, the study confined itself to grading, which students in law school experience only *after* they have evaluated a teacher, and did not assess the leniency-stringency variable in teacher behavior that students experience *before* they evaluate the teacher, e.g., instructor expectations regarding student attendance and preparation for class. The extent of pre-evaluation leniency seems likely to shape how a student fills out an evaluation questionnaire on an instructor, particularly when the student knows that faculty members in a program differ in leniency. In addition, the degree of leniency during a course can be expected to affect student mastery of material. Indeed, a plausible hypothesis is that lenient teacher standards prior to evaluation promotes favorable ratings from students even while depressing the test performance of the students. Unfortunately, a rigorous test of the hypothesis seems not to have been conducted.

For the reasons I have outlined, then, I do not believe we can yet be confident that we know what is being measured by student-completed teaching evaluation questionnaires or that the validity of the questionnaires is as high as the correlation coefficient (i.e., .44) found in the paper by Cashin. Any other conclusion seems to me to give insufficient weight to the serious limitations that characterize existing research on the questionnaires. As in the case of a drug that a pharmaceutical company seeks government approval to market, we should insist on a body of credible evidence that the questionnaires are safe and effective before we use them in personnel matters. Based on my review, such evidence does not presently appear to exist. In compiling and assessing this evidence, furthermore, we must recognize that the questionnaires seem to produce undesirable side effects—a matter to which I now turn.

THE CONSEQUENCES OF STUDENT RATINGS

We have, I believe, failed to appreciate the substantial *costs* of teaching evaluation questionnaires, and even if the questionnaires possess a degree of validity, I do not think we can afford to neglect these costs given the need of American universities to preserve the commitment of their faculty members and to promote the intellectual accomplishment of their students. An indication of the costs comes from a survey at a medium-size midwestern state university whose faculty members were asked a series of questions regarding their reaction to the adoption of student-completed teaching evaluation questionnaires [19]. The results of the survey are instructive and suggest the questionnaires damage educational quality and may therefore hinder students in maximizing their intellectual potential. Among the relevant survey findings are:

- Three-fourths of the faculty stated that, as a result of the questionnaires, their morale declined.
- The questionnaires alienated many faculty members from their university and from their students. The proportion of faculty members who, because of the questionnaires, reported a deterioration in the image they held of their university was substantially larger than the proportion who reported an improvement (46% versus 17%); the proportion of faculty members who reported diminished satisfaction with teaching was markedly higher than the

proportion who reported greater satisfaction (45% versus 12%); and the proportion of faculty members who reported an increase in the social distance between themselves and their students appreciably exceeded the proportion who reported a decrease (39% versus 19%).

- The questionnaires frequently reduced the academic standards that faculty members applied to students. While half to two-thirds reported no change in various requirements they had for students, the faculty members who altered their requirements because of the questionnaires were much more likely to decrease than to increase the amount and difficulty of material covered. Specifically, the amount of material covered was reduced by 22 percent of the faculty and expanded by just 7 percent, while the difficulty of course material was lowered by 38 percent of the faculty and raised by only 9 percent. Particularly notable was a finding on the rigor of course examinations: fully 33 percent of the faculty admitted that, as a result of the questionnaires, they lessened the difficulty of their tests; just 6 percent indicated they increased test difficulty.

CONCLUSION

In concluding, I have not read all of the research that has been done on the validity of student ratings, but given the studies I reviewed—which seem to be the most prominent studies on the subject—I believe there is good reason to be wary of the ratings. The problems of measurement, design, and data analysis I have considered are not minor imperfections, and they seriously undermine the argument that existing research has proven the ratings are reasonably valid measures of teaching effectiveness. Skepticism about the research is reinforced by a further problem—a problem that has been implicit in my discussion thus far. Specifically, studies to date appear to have generally employed just one indicator of student achievement, namely, performance on course examinations. Reliance on a single yardstick for student learning, and hence for teaching quality, is unfortunate and lessens the utility of the research that has been done on student ratings. Future work by social scientists on the validity of the ratings therefore needs to encompass "a diverse array of educational outcomes (e.g., taking advanced coursework, independent reading, problem solving skills, attitude changes)" in addition to test performance [16, p. 190].

A mistrust of student-completed teaching evaluation questionnaires should not, however, be grounded solely on reservations concerning their validity. Even if student ratings have some validity, we may lose more than we gain by relying on them in personnel decisions. Besides the costs of the questionnaires mentioned in the preceding section, let me suggest that it may not be coincidence the ratings and grade inflation entered university life at roughly the same point in time. To the extent the ratings encouraged grade inflation, they reduced the reward for student effort and skill and, in turn, probably eroded student motivation. The use of student ratings in decisions on faculty retention, promotion, and tenure may thus on balance be harming, not benefiting, the American system of higher education.

Finally, it appears that graduate-level programs in particular would be well-advised to assess, with rigorous social science research methods, the validity of the questionnaires they are using. As one established scholar in the field of teaching evaluation questionnaires observed, the focus of multisection validity studies to date has generally been on

> freshman or sophomore courses that emphasize lower-level learning, basic knowledge, and skills in a subject area. Teaching behaviors that best accomplish those learning outcomes may not work as well with higher-level outcomes such as critical thinking or synthesis. The relationship of student ratings to achievement, therefore, may not be as strong for teaching behaviors attempting to achieve higher-level outcomes [20, p. 63].

Courses that attempt to convey concepts and develop transposition skills, in short, are settings in which student-completed teaching-evaluation questionnaires are suspect. Among these courses are law school classes that rely on the Socratic method. At the very least, research should be undertaken on the validity of the questionnaires in such courses before student responses to the questionnaires are used in personnel decisions. Any other approach is unwise as a matter of personnel policy and may be indefensible as a matter of law.

* * *

Dr. Larry D. Barnett is Professor of Law at Widener University in Wilmington, Delaware. He has a dual background in social science and in law, holding both a Ph.D. and a J.D. Professor Barnett was the founder and editor of *Population Research and Policy Review.* He has also authored several legal texts.

ENDNOTES

1. For the teaching staff at public colleges and universities who are entitled by contract to continuing employment absent adequate cause, the due process guarantee appears to preclude the use of inaccurate indicators of instructional quality to justify the termination of employment. See *Perry v. Sindermann,* 408 U.S. 593, 603 (1972) (institutional policies and practices that have caused a terminated faculty member to expect continuing employment "obligate college officials to grant a hearing at his request, where he could be informed of the grounds for his nonretention and *challenge their sufficiency*" [emphasis supplied]). See generally John E. Nowak & Ronald D. Rotunda, *Constitutional Law* §§13.5(d), 13.8 (5th ed. 1995).

2. John Edward Davidson, Note, The Temptation of Performance Appraisal Abuse in Employment Litigation, 81 *Va. L. Rev.* 1605, 1614, 1628n.164 (1995).
3. See the first paragraph of the Conclusion, *infra* p. 344.
4. William E. Cashin "Student Ratings of Teaching: The Research Revisited" (Kansas State University Center for Faculty Evaluation & Development Idea Paper No. 32, 1995); Richard M. Felder, What Do They Know Anyway? 26 *Chemical Engineering Education* 134 (1992).
5. Given the nontechnical audience for this paper, I have adopted a nonstandard nomenclature. In social science, the effect variable and the explanatory variable are normally referred to as the dependent variable and the independent variable, respectively.
6. Upon finding a relationship, we will also want to know the sign and numerical value of the coefficient. The sign discloses the direction of a relationship, i.e., whether an increase in the explanatory variable causes an increase or a decrease in the effect variable. The sign of a correlation coefficient and the sign of a regression coefficient are interpreted in the same manner. The interpretation of the numerical value of a coefficient, however, depends on whether the coefficient is from correlation or from regression. The numerical value of a *correlation* coefficient specifies the *degree* to which quantitative change in the effect variable is associated with quantitative change in the explanatory variable. The numerical value of a *regression* coefficient specifies the *amount* of change in the effect variable that results from one unit of change in the explanatory variable.
7. William D. Berry & Stanley Feldman, *Multiple Regression in Practice*, Sage Publications, Newbury Park, CA, 1985 (italics in original). Copyright © 1985 by Sage Publications, Inc. Reprinted by permission of Sage Publications. It is possible for a coefficient to understate the actual relationship between an explanatory variable and the effect variable and, consequently, for the coefficient to become larger when an omitted variable is controlled. This would happen, for instance, where the omitted variable is related to the explanatory variable and to the effect variable, but the two relationships have coefficients with opposite signs: increases in the omitted variable would be raising the amount of one of the variables and lowering the amount of the other. See page 84-91 in Morris Rosenberg, The Logic of Survey Analysis, Basic Books, New York (1968). In social science research, however, the size of a coefficient is generally reduced by the introduction of a control variable. A reduction seems especially probable in the correlation coefficient that is claimed to exist between student evaluations of teaching quality and student achievement (see text *infra*), because the size of this coefficient exceeds what is normally found in the social sciences.
8. Cashin, *supra* endnote 4. Professor Cashin reproduces two sets of correlation coefficients, each from a different study. I will concentrate on just one of the studies, *infra* endnote 10, but the other study reports a coefficient of the same sign and comparable size (viz., .39) for the association of student evaluations of a teacher "overall" and student achievement. Kenneth A. Feldman, 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, 30,* pp. 583-645, 1989 (see p. 602). Both studies employed essentially the

same type of statistical analysis, however, and are therefore plagued by the same types of problems. These problems are discussed in the next section of the text.

9. Herbert W. Marsh, Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility, 76 *J. Educ. Psych.* 707 (1984).
10. The cited work is Peter A. Cohen, Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies, 51 *Rev. Educ. Res.* 281 (1981). For another methodological problem with this study, see note 17 *infra* and the accompanying text.
11. Cohen treated student judgments of "the amount and difficulty of the work the teacher expects of students" as one component of instructional quality. Whether it is an element of teaching quality or a separate factor, the amount and difficulty of work should be controlled under the conditions mentioned because it may explain much or all of the relationship detected between, on the one hand, evaluations of an instructor overall or on specific dimensions and, on the other, performance on examinations. From the studies he reviewed, Cohen calculated a negligible *mean* correlation coefficient for the relationship between the amount/difficulty of work and student achievement, but he also found a substantial range for the coefficients reported by the studies. Specifically, the interval for 95% of the coefficients extended from –.42 to +.39. *Id.,* at 293, 295. Individual studies may thus involve a nontrivial association between the perceived difficulty of teachers and the examination performance of students.
12. Cashin [4, p. 5] did not articulate the theoretical model he posited for the control, explanatory, and effect variables he considered, and the casual paths of the variables are thus not specified in his model. As indicated earlier in the text, I am treating the control variable as a common antecedent of both the explanatory variable and the effect variable. The control variable is not regarded as a factor that intervenes between the explanatory variable and the effect variable, nor is it considered to be the first variable in a three-variable causal chain.
13. Kenneth A. Feldman, The Association Between Student Ratings of Specific Instructional Dimensions and Student Achievement, 30 *Res. Higher Education,* 1989, p. 583.
14. Glen R. Sixbury and William E. Cashin, Description of Database for the IDEA Diagnostic Form 112,122-123 (Kansas State University Center for Faculty Evaluation and Development, Technical Report No. 9, 1995) (see the correlation coefficients between IDEA items 31, 33, and 35, on the one hand, and items 21, 22, and 23, on the other).
15. John A. Centra, Student Ratings of Instruction and Their Relationships to Student Learning, 14 *Am. Educ. Res. J.* 17, 21 (1977). In research on questionnaire validity, the measure of course difficulty and workload should not be confined to subjective data, i.e., to the views of students. Objective measures should be included as well, e.g., the number of pages of reading material assigned and the frequency with which instructors, in order to determine preparation for class, randomly question students about the material and penalize lack of preparation.
16. Herbert W. Marsh and Michael J. Dunkin, Students' Evaluations of University Teaching: A Multidimensional Perspective in *Higher Education: Handbook of Theory and Research* 143, 192 (John C. Smart, ed., Agathon Press, New York, Vol. 8, 1992).
17. Glenn Firebaugh, A Rule for Inferring Individual-Level Relationships from Aggregate Data, 43 *Am. Soc. Rev.* 557 (1978); John L. Hammond, Two Sources of Error in

Ecological Correlations, 38 *Am. Soc. Rev.* 764 (1973). "[A]n aggregate variable often measures a different construct than its namesake at the individual level. Often the aggregate-level variable taps more constructs than the individual-level variable" [Firebaugh [17], p. 560]. Where group-level data capture a factor that individual-level data do not and where this factor independently influences the effect variable, a coefficient computed with group-level data will be biased, i.e., will differ from the coefficient for individual-level data (Firebaugh [17]). In the context of student-completed teaching-evaluation questionnaires, the average rating of a teacher not only reflects the individual reaction of each student to the teacher but also shapes the general reputation of the teacher. Ratings as measured by section averages thus reflect in part the reputation of instructors, and if this reputation affects student performance on examinations (because, say, a reputation for being stern motivates students to prepare more fully than a reputation for being lax), a coefficient based on section averages is not a substitute for a coefficient based on individual-level data. The study by Cohen [10], also relies on group-level measures, and its results can be questioned for this reason.

18. John Palmer, Geoffrey Carliner, and Thomas Romer, Leniency, Learning, and Evaluations, 70 *J. Educ. Psych.* 855, 862 (1978) (emphasis supplied). Student ratings may be valid, as Felder stated, where the ratings that students give a teacher improve with the grades received by the students: "if students learn more from a teacher . . . , one would expect both their grades and their ratings to be higher [4, p. 134]. However, research reporting that favorable ratings for instructors are associated with higher grades for students must control factors that could have produced the association. One of these factors may be the difference between teachers in the leniency of their course requirements: teachers who are generally lenient may be rewarded by students with better ratings than teachers who are generally demanding; in addition, lenient teachers are by definition likely to assign elevated letter grades. Another factor that may require control is the level of interest that students bring to the subject matter of a course, since the rating of a teacher by a student and the grade of the student may rise with the prior interest of the student in the topic. See Kenneth A. Feldman, Grades and College Students' Evaluations of their Courses and Teachers, 4 *Res. Higher Educ.* 69, 86-91 (1976). In these circumstances, a spurious relationship will exist between student ratings and grades. Studies to eliminate the possibility of a spurious relationship should employ the multisection design described earlier in the text; relationships between variables should be estimated with individual-level data; and statistical analyses should include controls *inter alia* for the difference between instructors in leniency and for initial student interest in the topic. Additionally, numerical scores on the (common) examination should be the measure of student achievement; letter grades are inadvisable measure if the instructor for each section possesses (and exercises) the discretion to assign grades to students in the section.
19. James J. Ryan, James A. Anderson, and Allen B. Birchler, Student Evaluation: The Faculty Responds, 12 *Res. Higher Educ.* 317 (1980). Tables 2 and 3 of this article are the source of the figures I use here.
20. John A. Centra, *Reflective Faculty Evaluation,* Jossey-Bass Publishers, San Francisco, (1993). If the hypothesis advanced by Centra is correct, the level of validity may differ from one specialty to another within a single discipline. Specifically, student

assessments of teaching in a discipline are likely to be less accurate when the subject matter is indefinite and unstructured than when the subject matter is precise and highly organized.

Direct reprint requests to:

Larry D. Barnett
School of Law
Widener University
P.O. Box 7474
Wilmington, DE 19803-0474