Reliability and Validity of Student Evaluations: Testing Models Versus Survey Research Models*

Hugh Hinton, *The University of Toledo*

Student evaluations of faculty teaching have four recognized functions. They provide diagnostic feedback for faculty, an evaluative tool for personnel decisions, information for students, and a subject for academic research. Regardless of how the evaluations are used, the user must be concerned about the reliability of the evaluation instrument and the validity of the student responses. Faculty members have a natural self-interest in the reliability and validity of student evaluations when used as an evaluative tool for personnel decisions.

Much of the existing literature on reliability and validity conceptualizes student evaluations as a "test," whose reliability is to be estimated and whose validation is to be determined. Student evaluations, however, resemble public opinion surveys more than they do objective tests. Consequently, an alternative approach is to conceptualize student evaluations as survey research rather than as tests. Such a conceptual framework provides new insights into student evaluations and an entirely different dimension to the question of reliability and validity.

Personnel systems in American colleges and universities, as in other organizations, must solve the problem of allocating organizational resources to reward and reinforce productive behavior. They must grapple with the universal problem of defining, measuring, and rewarding merit. However, academic personnel systems differ in that they do not share a universally accepted "model" of who is to evaluate merit and how it is to be done. Instead, evaluation in higher education generally uses mixtures of three models for allocating rewards, two of which are commonly found in other organizations.

The most common model for evaluating merit is the "supervisor-subordinate" model, in which the performance of a member is appraised by a supervisor/superior. Although there are a number of different approaches and instruments (i.e., trait-oriented or behavior-oriented, comparison or forced choice), the defining characteristic is compatibility with the formal organizational hierarchy. Most textbook treatments of performance appraisals restrict themselves almost exclusively to this model, and a large body of normative literature exists in human resource management and empirical research in organization theory focusing on this model.[1] All universities use this model to the extent that university administrators are involved in personnel decisions allocating organizational rewards.

The second common model for evaluating merit rejects the hierarchical framework in favor of some variation of peer evaluation. More commonly known as the "professional model,"[2] it is based on the premise that the performance of members of certain professions can only be adequately evaluated by other like professionals.[3] Peer evaluation, self-governance, and tenure are central to the principles of the American Association of University Professors, and are found to some extent in the accreditation requirements of numerous accrediting bodies.

University personnel systems combine some of the supervisor-subordinate and the professional models of performance evaluation. However, the third model of evaluation is not only common to universities but is used in a way that is virtually unique to them—the 'clientele" or the "service recipient" model. In this model, clientele or service recipients have a significant role in the evaluation process. The form this model takes in universities is, of course, using student evaluations in decisions to hire, compensate, retain, promote, and tenure academic personnel.

This model can also be seen in the private sector, where customer feedback can be a significant factor in personnel decisions. However, in the public and nonprofit sectors, use of this model is limited. For example, even though police and social work agencies frequently have procedures for handling comments and complaints by citizen "clients," generally only those complaints that have been substantiated have any effect on personnel decisions.

Academic personnel systems, on the other hand, not only accept unsubstantiated subordinate (student) comments and complaints, they actually solicit them. Traditionally, a central component in this practice has been a formal teacher evaluation instrument to be completed by students.

It is not known precisely how widely teaching evaluations are used or how much importance is attributed to them. On the one hand, one recent study found that the percentage of private liberal arts colleges that used student opinion in evaluating teaching performance increased from 29% in 1973 to 67% in 1983 (Seldin 1984, 48). On the other hand, assertions such as "most faculty members in most universities have been granted tenure in spite of their performance in the classroom" (Baker 1990) are widespread.

An enormous volume of academic literature on this subject exists, and faculty have applied their own evaluative research techniques to the criticisms that students have made of them. According to one source, there are 1,300 citations in the Educational Resources Information Center data base under "student evaluation of teacher performance" at the post-secondary level (Cashin 1990). This faculty interest suggests that student evaluations are playing an increasingly important role in defining meritor-

https://doi.org/10.2307/420006 Published online by Cambridge University Press

ious teaching and allocating organizational rewards.[4]

The Controversy

Studies of student evaluations of faculty teaching fall into two broad categories. Most, using a variety of methodological approaches, deal with the question of student characteristics that are closely associated with student ratings of faculty. Only a few of these examine, even casually, the question of the validity and reliability of student evaluations. These include eclectic methodological approaches, covering the range from impressionistic/anecdotal/experiential observations to sophisticated statistical modeling. Louis Goldman (1990; also see Rutland 1990, 1-2), writing in *The Chronicle of Higher Education,* exemplifies the former:

> Student evaluations tell us much more about the students than about the instructors or courses they are evaluating. Several variables contribute to the subjectivity of students' responses, the most dominant ones being the interests, needs, and background of each student.

These assertions provoked a storm of protest from dissenters, filling the "Letters to the Editor" section the following month.[5] One writer asserts, "If there is anything the research (Feldman 1976; Marsh 1984) is agreed upon, it is that student ratings are statistically reliable" (Cashin). An examination of the works of these and other researchers, however, does not reveal such an unequivocal conclusion. After reviewing 72 prior studies, Kenneth Feldman cautions that

> . . . the question can still be . . . raised as to whether students are in a position to make accurate judgments about certain matters, including the instructor's degree of knowledge of the subject matter of the course, the instructor's preparation and organization of the course, and the instructor's ability to explain clearly (1976b, 266).

Herbert Marsh, after reviewing 136 studies, states in the abstract that ". . . class average student ratings are . . . reliable and stable; . . . relatively valid. . . ; [and] . . . relatively unaffected by . . . potential biases" (1984, 707). However, in the article itself, he is far more cautious and states:

> Research . . . findings also demonstrate that student ratings may have some halo effect, have at least some unreliability, have only modest agreement with some criteria of effective teaching, and are probably affected by some potential sources of biases. . . (1984, 749).

Testing Model

The literature on teaching evaluations is primarily the domain of faculty in Colleges of Education (although Feldman is a sociologist). For example, of the 136 studies reviewed by Marsh, all but 10 have been published in education journals or books, or presented at education association conferences. These researchers, in both their analytic and validation research, use methodologies and assumptions drawn from their familiar domain of educational testing and measurement. There is a tendency to unquestioningly accept the philosophical and theoretical dimensions underlying these studies, particularly the unstated assumption that the reliability and validity of student evaluations can be determined by applying the same criteria used for tests. A close and careful reading of this literature reveals some inconsistencies that might be traced to assuming that student evaluations are another form of a "test," and indicates that the conclusions are not nearly as clear as might appear at a casual glance.

Reliability

In social research, reliability refers to the dependability or the "relative absence of errors of measurement in a measuring instrument" (Kerlinger 1986, 430). Kenneth Feldman (1977) has provided an excellent attempt to grapple with the problem of the reliability of student evaluations, using traditional test reliability theory from educational psychology. In classical reliability theory, according to Feldman,

> observed scores of individuals on psychological tests are conceived as comprising some "true" component, which is variously defined. . . , plus an "error" component. Reliability is then defined as the ratio of the variance in true scores to the variance in observed scores (1977, 224).[6]

Feldman summarizes the approaches used to estimate reliability —test-retest, parallel test, split-half, and internal consistency—and opts to apply the latter two to determine the reliability of student ratings of faculty. In his review of other research that has attempted to estimate the reliability of student ratings by determining the consistency among these ratings, he identifies six different procedures that have been used (1977, 225-28):

1. calculating the product-moment correlation of their ratings;
2. correlating the ratings of randomly drawn pairs of raters;
3. using the coefficient of interclass correlation;
4. using the generalized reliability formula developed by Horst;
5. determining two mean scores gained by dividing each class into two subgroups of students;
6. applying generalizability theory, allowing for a multidimensional interpretation of error.

Feldman, however, does not question whether assumptions and approaches designed to estimate the reliability of a personality test are appropriate to estimating the reliability of student evaluations. He and other researchers attempting to estimate reliability using this assumption have uncovered inconsistencies that are traceable to their methodology. Sometimes these are ignored, and at others, the methodology is modified to fit the data. Feldman, for example, encounters these kinds of problems several times and resolves them primarily by ignoring them. He states:

> The assumption of random sampling of students is problematic . . . since students in part self-select themselves into courses. . . . Thus the assumption of random sampling is often relaxed by introducing the idea of an unspecified population of students "like those observed" (1978, 200).

Further, Feldman states that "raters are considered as functioning very much as 'items' do in conventional tests" (1977, 231). However,

he further states, "Ratings are justifiably averaged or polled if they have been made independently"; then admits that student evaluations are likely contaminated by being highly collaborative (1977, 231). He further admits:

> The teacher rating situation is different because the "object" to be rated (the teacher, the course, or both) is the very entity about which students have been, in part, mutually influencing one another's opinions and jointly forming their assessments (1977, 232).

Feldman concedes "that students are only moderately consistent in rating their teachers, and . . . the . . . variability . . . is associated with various student experiences and attributes" (1977, 248). In other words, he concedes that student evaluations do not meet the criteria of interrater reliability that he has set. Turning to the question of student objectivity, Feldman reached no firm conclusion, but did make the startling assertion that "there is little direct and systematic evidence . . . that an increase in the objectivity of ratings brings about an increase in interrater reliability" (1977, 251). Finally, he cites research that argues "that student characteristics that are associated with ratings need not be regarded as biasing elements," and that "some degree of inconsistency among students in their evaluation of teachers is considered reasonable. . ." (1977, 253). He concludes his reliability attempt somewhat lamely by interpreting his findings to mean that the various correlates of within-class ratings are "biasing" elements if the student ratings are claimed to be objective descriptions, or "natural" influences on social perception if they are intended to measure the subjective reactions of students to teachers and courses (1977, 257-58).

On the other hand, Herbert Marsh (1984, 716) argues against using item analysis and interrater agreement to determine reliability, because

> the internal consistency among items is consistently high, but it provides an inflated estimate of reliability because it ignores the substantial portion of error due to the lack of agreement among different students, and so it generally should not be used. . . .

However, Marsh does argue that, with "a sufficient number of students, the reliability of class-average student ratings compares favorably with that of the best objective tests" (1984, 717).

Validity

Determining validity, on the other hand, is less technical and more philosophical than determining reliability. Validity—or operational validity, to be technically correct—refers to the question of "what" is being measured, and whether the instrument measures what we want to measure (Kerlinger 1986, 444-45). In the words of Fred Kerlinger:

> Achieving reliability is to a large extent a technical matter. Validity, however, is much more than a technique. It bores into the essence of science itself. It also bores into philosophy (1986, 459).

Faculty often view the validity of results of student evaluations with skepticism and question the motives and criteria students use in evaluating faculty. Two researchers, for example, observed:

> From the very beginning of their use, faculty have expressed reservations about the meaning (validity) of student responses regarding teaching effectiveness. Put simply, faculty have argued that they and students use different criteria in evaluating teaching. Naturally, faculty view their own standards as being more relevant for, or consistent with, the long-run mission of higher education (Baum and Brown 1980, 234).

Thus far, research has not presented sufficiently unequivocal evidence of validity to quiet these reservations. Richard Miller, for example, notes:

> The literature on the validity of student evaluations of classroom teaching is extensive and conflicting, and sometimes it does not rise above conceptual and methodological mediocrity (1987, 54).

There are four normally recognized types of validity—predictive, concurrent, content, and construct (see Kerlinger 1986, 444-54), although some texts vary these types and their definitions (e.g., O'Sullivan and Rassel 1989, 91-98; or Welch and Comer 1988, 42-45). The two types most commonly employed to validate student evaluations are content and construct. Content validation refers to the "*representativeness* or *sampling adequacy* of the content," and asks the question: "Is the substance or content of this measure representative of the content or the universe of content of the property being measured?" (Kerlinger 1986, 445-46). Construct validity "unites psychometric notions with scientific theoretical notions" (Kerlinger 1986, 448). In the words of Herbert Marsh:

> The construct validity of students' evaluations requires that they be related to variables that are indicative of effective teaching, but relatively uncorrelated with variables that are not (i.e., potential biases) (1984, 730).

In several excellent articles, Marsh provides one of the most thorough attempts to validate student evaluations, primarily using the construct validation approach (1977, 1980, 1982a, 1982b, 1984; Marsh, Fleiner, and Thomas 1975; Marsh and Overall 1980; Marsh, Overall, and Kesler 1979). Marsh accepts the premise that prior research findings have been contradictory, but defends the construct validation approach as justified by four underlying perspectives (1984, 708-09):

1. Teaching effectiveness is multifaceted.
2. Since there is no single criteria of effective teaching, ratings must be shown as related to a variety of other indicators of effective teaching.
3. Different dimensions or factors of the ratings must be significantly correlated with criteria to which it is logically and theoretically related.
4. Bias interpretations must be made in the context of an explicit definition of what constitutes a bias.

Marsh subsequently posits several accepted criteria of effective teaching (1984, 720):

> student learning, changes in student behaviors, instructor self-evaluations, peer/administrator evaluations, behaviors observed by trained observers, and effects of experimental manipulations.

Construct validation studies have had some difficulty in demonstrating a relationship between student ratings and some accepted characteristics of good teaching. For one, attempts to validate student evaluations through comparison with faculty self-evaluations have returned mixed results. Marsh, for example, concluded that existing studies demonstrated that "the fact that students' evaluations show significant agreement with instructor self-evaluations provides a demonstration of their validity. . ." (1984, 723). Feldman, however, has not been as unambiguous. After reviewing available studies of students' reported desired characteristics of effective teaching, he found that the dimensions most consistently highly associated were "stimulation of interest and clarity," "knowledge of subject matter," class preparation, and enthusiasm (1976b, 263)—hardly exhausting all the factors that would rate as important. Not unexpectedly, Feldman concludes that

> the question can still be . . . raised as to whether students are in a position to make accurate judgements about certain matters, including the instructor's degree of knowledge of the subject matter of the course, the instructor's preparation and organization of the course, and the instructor's ability to explain clearly (1976b, 266).

In a later review of 31 studies comparing the differential importance of various components of teaching between faculty and students, he discovered an average correlation of agreement of +.71, with a combined Z of +19.421 and a p < .001 (1988, 298). However, there is sufficient variation in the findings of these studies to raise some question about external validity. One study, for example, found the student-faculty correlations high (and statistically significant) for the social sciences, humanities, and engineering, but an insignificant correlation in the natural sciences (Marques, Lane, and Dorfman 1979). Other studies found positive but statistically insignificant correlations in business schools (Baum and Brown 1980) and negative correlations in a sociology department (Norr and Crittenden 1975).

Also, Feldman finds notable discrepancies in four important dimensions. Students reported that three of these dimensions were of low or moderate importance (clarity of course requirements, outcome of instruction, intellectual challenge), yet they all correlate highly with student evaluations of faculty teaching. The fourth dimension (teacher's intelligence) correlates weakly with overall evaluations, but was reported to be of moderate importance by students (1988, 319).

Construct validation seeking a relationship between student ratings and faculty research productivity has been relatively unsuccessful, as have studies using comparisons between students and peer/supervisor ratings. In the latter case, studies have reached a number of inconclusive findings. Among these are that there is a general lack of consensus among peers when rating colleagues (Centra 1975) and that there is generally significant correlations between peer and supervisor ratings, but that these were unrelated to student ratings (Morsh, Burgess, and Smith 1956). Also, a relationship between student ratings and faculty self-evaluations has been discovered, but neither correlated with peer/supervisor ratings (Webb and Nolan 1955). Literature reviews have failed to uncover any studies that support the validity of peer evaluations (e.g., Marsh 1984; Centra 1979; and French-Lazovich 1981), although one review somewhat defensively asserts that "these findings neither support nor refute the validity of student ratings" (Marsh 1984, 725). Another, however, even argues for the superiority of student evaluations over peer ratings because the latter are "(1) less sensitive, reliable, and valid; (2) more threatening and disruptive of faculty morale; and (3) more affected by noninstructional factors such as research productivity" (Murray 1980, 45; cited in Marsh 1984). The author fails to note that research has also demonstrated that student ratings of faculty are also affected by noninstructional factors.

On the other hand, research has been more successful in identifying a relationship between the observations of faculty by external observers, on the one hand, and student ratings of faculty, on the other. One research review, for example, concluded that student evaluations "can be accurately predicted from outside observer reports of specific classroom teaching behaviors" (Murray 1980, 31; cited in Marsh 1984).

To validate student ratings using the construct validation approach, however, requires that these ratings not be related to factors not seen to be part of good teaching. Researchers have encountered considerable difficulty when attempting to dismiss either a nonexistent or negative relationship with an accepted characteristic, or a positive relationship with a negative or unrelated characteristic. Herbert Marsh, after reviewing a number of his and other studies concerning these "potential biases" concluded that "between 5% and 20% of the variance in student ratings can be explained" by background factors, primarily prior subject interest, expected grades, and perhaps course workload/difficulty (1984, 731). However, he then makes an interesting assertion:

> The finding that a set of background characteristics are correlated with students' evaluations of teaching effectiveness should not be interpreted to mean that the ratings are biased, although this conclusion is often inferred by researchers (Marsh 1984, 731).

In his subsequent review and analysis of bias research of four background variables (the above three plus class size), Marsh discounts each of these as biasing factors. Other reviewers, however, are not so certain. Kenneth Feldman, after reviewing 58 studies of the relationship between class size and student ratings, found that class size was associated with the interpersonal and facilitative rating dimensions. However, for most other rating dimensions, he found that "global ratings of teachers are as likely to be inversely associated with class size as not to be related at all" (1984, 71-72). However, Feldman uses this association to validate student ratings.

> It clearly makes sense that class size has been found . . . to be related . . . to those instructional dimensions involving teachers' interactions and

> interrelationships with students (encouraging class discussion, giving feedback to students, being fair to students when evaluating their work, caring about and respecting them, being open to their opinions, and available to help them) (1984, 77).

After all this tortuous reasoning, however, Feldman reaches a conclusion that raises some questions about the underlying premises of the construct validation approach.

> Yet the matter is more complex. Even though the particular patterning of the differential relationships of class size with different instructional dimensions that has been found is logical, this does not preclude the possibility of bias in the evaluations (1984, 78).

Marsh, also, makes a number of interesting observations that imply that he is not as certain of his conclusion as he suggests. Many of his own studies attempt to demolish the "grading leniency hypothesis," which argues that faculty giving higher grades are rewarded with higher evaluations. He admits that the relationship between grades and evaluations is positive, but argues that it still might not be an invalidating factor. He posits three possible hypotheses to explain this positive relationship. Higher grades might be due to more effective teaching, increased student satisfaction with higher grades, or prior student characteristics, such as student interests. He argues, in effect, that this positive relationship is a spurious correlation, and he reviews his and other research that dismisses grades as an evaluation bias.

In one study (with Overall 1979), Marsh concluded that his data argued "against the interpretation of the expected grade effect as a bias." However, he later concedes that

> the fact that expected grades were more positively correlated with student ratings than with faculty self-evaluations may provide some support for a grading leniency bias (Marsh 1984, 739).

Marsh's later comment about another of his studies which questioned the grading leniency hypothesis (1982c) in essence questioned the external validity of his own study by stating that

> this study was in a setting where differences due to grading leniency were minimized, there was no basis for contending that the grading leniency effect does not operate in other situations (1984, 739).

And his comments about two other of his studies that support the validity hypothesis (Marsh, Fleiner, and Thomas 1975; Marsh and Overall 1980) is instructive, for he stated that

> support for the validity hypothesis found here does not deny the appropriateness of other interpretations in other situations (1984, 740).

And he concludes his review about expected grades as a biasing factor by stating that "Evidence . . . does not rule out the possibility that a grading leniency effect operates simultaneously" (1984, 741).

Other reviews are no less ambiguous about this grading leniency hypothesis. In one of the most extensive meta-analyses of this subject, Kenneth Feldman argued that

> under certain conditions (and for certain classes), expected or actual grade is indeed related to evaluation, even strongly so, whereas for other conditions and classes there is little or no relationship (and possibly even a negative one) (1976a, 83).

Reliability and Validity Reexamined

Research about the reliability and validity of student ratings of faculty is far less conclusive than superficial reviews might assert. Although charges of research mediocrity should not be ipso facto dismissed, there are other alternative reasons for the ambiguity and inconclusiveness in this research. In particular, objective tests may be sufficiently different so that they are not an appropriate model for an analysis of student ratings. To determine this, it is important to examine the fundamental characteristics of tests and distinguish these from student ratings.

All objective tests, whether they be achievement, intelligence, attitudinal, or personality, share several commonalities. First, they attempt to achieve objectivity by reducing observer and judgmental variances to zero. Second, they are premised on the possibility of objectifying the subjective. And third, they are measures of variables, from which characteristics of individuals can be inferred.[7] In sum, tests are generally devised to measure characteristics about the person taking the test (the testee), or more accurately, to allow us to infer certain characteristics, based on the testee's response to test items. They are "focused mainly on measures responded to by the subject being measured" (Kerlinger 1986, 514). With student evaluations, in contrast, the "test" (evaluation) is taken to reveal characteristics of a third party (i.e., faculty) rather than of the person taking the "test."

Using objective tests as a model for determining the reliability and validity of student evaluations of faculty thus has major conceptual flaws. Tests and student evaluations measure what can be designated as different "realities," a fact generally ignored in the literature. A test measures characteristics of the testee, whereas student ratings measure a student's *perception* of a faculty member. Probably because of the influence of the testing model on student evaluation studies, researchers at times assume objectivity where it should rather be proven. In particular, there are numerous examples where researchers accepted uncritically the self-reported results of student evaluations. P. A. Cohen, for example, reviewed 68 multisection validity studies, and found that student achievement consistently correlated with student ratings of skill, overall course, structure, student progress, and overall instructor (Cohen 1981). He makes the questionable assumption that *reported* achievement was *actual* achievement. More precisely, the unanswered question was whether students' *perceived* learning equaled their actual learning. In contrast, Feldman recognized the difference between learning as measured by "objective" measures and students' own perceptions of their learning. The former is "related to ratings at about the same strength . . . as that found for the association between grades and ratings," whereas the latter "is associated much more strongly with course and teacher ratings" (1977, 236).

One solution to this conceptual

incompatibility, while still maintaining the framework of testing theory, would be to characterize students evaluating faculty as equivalent to a trained observer rating third-party behavior. Observer ratings have been characterized as "measures of individuals and their reactions, characteristics, and behaviors by observers. The contrast, then, is between the subject as he sees himself and the subject as others see him" (Feldman 1977, 236).

Even viewing student evaluations as observations of behavior rather than as objective tests, however, still presents some serious problems. In the first place, most reliability and validity studies of student evaluations do not draw such a clear distinction, and generally imply—sometimes explicitly—a treatment of evaluations as a "test." Second, the student observer, as part of the measuring instrument, "must digest the information derived from his observations and then make inferences about the constructs" (Kerlinger 1986, 505). In objective tests, on the other hand, the (presumably trained) social scientist, rather than the testee, draws the inferences. Third, validation of student evaluations as observations is problematic. Although construct validation is generally recognized as the appropriate validation approach, it is handicapped by the lack of a universally acceptable model of "good teaching." And fourth, reliability estimation is an even more serious problem, since the most usual definition is "the agreement among observers" (Kerlinger 1986, 507). This agreement is extremely problematical and conceptually perverse when we conceive of students as observers.

It is therefore questionable whether it is appropriate to view student evaluations as a type of objective test or an observer rating and, thus, susceptible to the same validation standards as tests. The difficulties encountered in attempts to make student evaluations conform to the standards of tests may provide a clue that this is not an appropriate analogy. Instead of assuming that student evaluations are a test, we should perhaps concede that they are what they appear to be—the opinion of students about faculty and courses. Just as a president's approval rating may tell us more about the public than about him, perhaps student ratings of faculty provide more information about students than about faculty. Instead of the testing analogy drawn from education and psychology, perhaps the public opinion analogy from the social sciences is worth exploring. The methodology of survey research, as familiar to social scientists as classical test validation theory is to education faculty, can provide a different, and possibly more fruitful, approach to studying student evaluations of faculty.

Survey Research Model

Survey research is that branch of social science research that studies the characteristics, attitudes, values, and behaviors of populations.[8] It has been characterized as "the best method available . . . for describing a population too large to observe directly" (Babbie 1986, 203-04). Backstrom and Hursh are more precise in their contention that

> certain kinds of knowledge can best be obtained by survey techniques. Generalizations about the characteristics of, or predictions about, the behavior of a great body of people require measurements along a broad spectrum of opinions, attitudes, feelings, beliefs, ideals, information, and understanding (1981, 8-9).

The survey research model therefore begins with different basic assumptions and objectives than the testing model. First, it does not study one group to learn the characteristics of another group. An analysis of student ratings of faculty would be approached as a source of knowledge about students. If we want knowledge about the characteristics and quality of faculty teaching, we should study a random sample of faculty. Second, because the respondent is also the subject of study, it abandons the pretext that students are objective observers about a third party. A response is presumed to be at least somewhat subjective, but provides a basis for inferring certain characteristics. And third, and possibly most important for our purposes, criteria for estimating the reliability and determining the validity for survey research is different than for objective tests.

In survey research, the questions of validity and reliability center around choosing the sample and constructing the survey instrument (see Babbie 1986, ch. 9). Since survey research generally takes samples of the universe under study, the validity of the survey instrument is checked against some outside criterion representative of this universe—such as a census (Babbie 1986, ch. 9). Survey research questionnaires are designed to include factual data that can be checked. In the words of Kerlinger:

> the reliability of personal factual items . . . is high. The reliability of attitude responses is harder to determine, because a change in response may signify a real change of attitude. The reliability of average responses is higher than the reliability of individual responses (1967, 401; also see Parten 1950, 496-98).

In contrast, Babbie contends that survey research is "generally weak on validity and strong on reliability" because survey responses are only approximate, rather than precise, indicators of what the instrument designers intended. Reliability, he contends, is a different matter, by presenting all subjects with a "standardized stimulus" and wording the questions carefully (1986, 232).

On the other hand, survey research does have some identifiable weaknesses. Kerlinger identifies two: it does not penetrate very deeply, emphasizing the scope of the information at the expense of depth; and it is demanding of time and money (1986, 407). The latter concern probably does not apply to student evaluations, since they are typically self-administered by a captive population.

Even if we concede that student ratings provide more knowledge about students than about faculty, there is still a problem to resolve. We have already questioned the premise that a class is a representative sample of the universe of students, but we can resolve this dilemma if we define the universe as that particular class and concede a lack of external validity for student evaluations. Indeed, there is at least some evidence to justify such a concession. Or, perhaps, taking a random sample of the

national universe of students might be an appropriate solution.

Conclusion

Although space has precluded an exhaustive examination of all the varied procedures for estimating reliability and the types of validity, there is sufficient evidence to suggest caution when using student ratings of faculty in personnel decisions and in allocating organizational resources. The concept of student ratings as either an objective test or as trained observers does not unequivocally meet the necessary standards for reliability and validity. When used to provide information about faculty, they should be recognized for what they are—student perceptions rather than objective facts. Consequently, when used in personnel decisions, they should be used with the justifiable caution that a number of faculty committees have recommended. At a major midwestern university, for example, a committee report cautioned that

> . . . we believe that student evaluation of teaching should not be the only evidence of teaching effectiveness reviewed in a faculty personnel decision. Teaching is a complex phenomenon, one that should not be reduced to a single rating form or to opinion garnered from a single source (Student Evaluation of Teaching Advisory Committee 1986).

And a similar committee at another midwestern university recommended that:

> All Departments . . . must provide the Dean with evidence of each full-time faculty member's teaching performance, including the results of student teaching evaluations. . . . Because evidence drawn from a variety of sources is more reliable and valid than evidence drawn from only one source, we urge that student teaching evaluations be used in conjunction with some of the following types of appropriate evidence. . . (The University of Toledo 1987).

On the other hand, student ratings can be viewed as a reliable and valid form of survey research, providing us with a wealth of knowledge about the attitudes, behavior, characteristics, and values of students. If we want knowledge about faculty, perhaps we should study faculty.

Notes

*I would like to acknowledge my colleagues Professors Lynn Bachelor and David Wilson, who provided assistance and advice for this paper.

1. In the classic public personnel administration texts of O. Glen Stahl and the Nigro's, it is explicitly assumed that employee evaluations were done by superiors/supervisors. Even more recent texts omit other models of evaluation. See for example George L. Morrisey (1983), Nicholas P. Lovrich (1983), or Wilber Rich (1989).
2. The literature on professionalism in public service is enormous and varied. The standard sources include Mosher (1982), Heclo (1977), Benveniste (1977), and Abrahamson (1967).
3. Literature on professionalization, however, does not always include peer evaluation as a necessary component of the professional model. Many police reformers appear to equate professionalization with depoliticization and bureaucratization that is more compatible with the supervisor-subordinate model than with the peer review model. This theme underlies the work of one of the fathers of police professionalization, O. W. Wilson. See Wilson (1972).
4. There are several excellent overviews and reviews of this literature. The most exhaustive are the 12 reviews by Kenneth Feldman published in *Research in Higher Education* from 1976 to 1988. More manageable are Marsh (1984) or Miller (1987).
5. September 5, 1990. Of the seven letters published in response to this article, only one could be said to have been supportive.
6. For a detailed discussion of classical reliability theory, see Lord and Novick (1968, ch. 2) or Wiggins (1973, ch. 7).
7. There are a number of excellent discussions of objective tests. One of the classic treatments is Kerlinger 1986, ch. 27.
8. The literature on survey research is extensive. Among the classic treatments are Kerlinger 1986, ch. 22; Backstrom and Hurch 1981; and Babbie 1986, ch. 9.

References

Abrahamson, Mark, ed. 1967. *The Professional in the Organization.* Chicago: Rand McNally.

Babbie, Earl. 1986. *The Practice of Social Research,* 4th ed. Belmont: Wadsworth Publishing Co.

Backstrom, Charles H., and Gerald D. Hursh. 1981. *Survey Research,* 2nd ed. Evanston: Northwestern University Press.

Baker, Steven J. 1990. "Letters to the Editor." *The Chronicle of Higher Education* September 5: B3.

Baum, P., and W. W. Brown. 1980. "Student and Faculty Perceptions of Teaching Effectiveness." *Research in Higher Education* 13: 233-42.

Benveniste, Guy. 1977. *Politics of Expertise,* 2nd ed. San Francisco: Boyd and Fraser.

Cashin, William E. 1990. "Letters to the Editor." *The Chronicle of Higher Education* September 5: B4.

Centra, J. A. 1975. "Colleagues as Raters of Classroom Instruction." *Journal of Higher Education* 46: 327-37.

Centra, J. A. 1979. *Determining Faculty Effectiveness.* San Francisco: Jossey-Bass.

Cohen, P. A. 1981. "Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies." *Review of Educational Research* 51: 281-309.

Feldman, Kenneth. 1976a. "Grades and College Students' Evaluations of Their Courses and Teachers." *Research in Higher Education* 4: 69-111.

Feldman, Kenneth. 1976b. "The Superior College Teacher From the Students' View." *Research in Higher Education* 5: 243-88.

Feldman, Kenneth. 1977. "Consistency and Variability Among College Students in Rating Their Teachers and Courses: A Review and Analysis." *Research in Higher Education* 6: 223-74.

Feldman, Kenneth. 1978. "Course Characteristics and College Students' Ratings of Their Teachers: What We Know and What We Don't." *Research in Higher Education* 5: 199-242.

Feldman, Kenneth. 1984. "Class Size and College Students' Evaluations of Teachers and Courses: A Closer Look." *Research in Higher Education* 21: 45-91.

Feldman, Kenneth. 1988. "Effective College Teaching From the Students' and Faculty's View: Matched or Mismatched Priorities?" *Research in Higher Education* 28: 291-328.

French-Lazovich, G. 1981. "Peer Reviews: Documentary Evidence in the Evaluation of Teaching." In *Handbook of Teacher Evaluation,* ed. J. Millman. Beverly Hills, CA: Sage.

Goldman, Louis. 1990. "Student Evaluations of Their Professors Rarely Provide a Fair Measure of Teaching Ability." *The Chronicle of Higher Education* September 5: B2.

Heclo, Hugh. 1977. *A Government of Strangers.* Washington, DC: The Brookings Institution.

Kerlinger, Fred N. 1986. *Foundations of Behavioral Research: Educational and Psychological Inquiry,* 3rd ed. New York: Holt, Rinehard and Winston.

Lord, F. M., and M. R. Novick. 1968. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Lovrich, Nicholas P., Jr. 1983. "Assessing the Performance of the Individual on the Job." In *Handbook on Public Personnel Administration and Labor Relations,* ed. Jack Rabin et al. New York: Marcel Dekker.

Marques, T. E., D. M. Lane, and P. W. Dorfman. 1979. "Toward the Development of a System for Instructional Evaluation: Is There Consensus Regarding What Constitutes Effective Teaching?" *Journal of Educational Psychology* 71: 840-49.

Marsh, Herbert. 1977. "The Validity of

Students' Evaluations: Classroom Evaluations of Instructors Independently Nominated as Best and Worst Teachers by Graduating Seniors." *American Educational Research Journal* 14: 441-47.
Marsh, Herbert. 1980. "Research on Students' Evaluations of Teaching Effectiveness." *Instructional Evaluation* 4: 5-13.
Marsh, Herbert. 1982a. "SSEQ: A Reliable, Valid, and Useful Instrument for Collecting Students' Evaluations of University Teaching." *British Journal of Educational Psychology* 52: 77-95.
Marsh, Herbert. 1982b. "Validity of Students' Evaluations of College Teaching: A Multitrait-Multimethod Analysis." *Journal of Educational Psychology* 74: 264-79.
Marsh, Herbert. 1982c. "Factors Affecting Students' Evaluations of the Same Course Taught by the Same Instructor on Different Occasions." *American Educational Research Journal* 19: 485-97.
Marsh, Herbert. 1984. "Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility." *Journal of Higher Education* 76: 707-54.
Marsh, Herbert, H. Fleiner, and C. S. Thomas. 1975. "Validity and Usefulness of Student Evaluations of Instructional Quality." *Journal of Educational Psychology* 67: 833-39.
Marsh, Herbert, and J. U. Overall. 1979. "Validity of Students' Evaluations of Teaching: A Comparison with Instructor Self-Evaluations by Teaching Assistants, Undergraduate Faculty, and Graduate Faculty." Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
Marsh, Herbert, and J. U. Overall. 1980. "Validity of Students' Evaluations of Teaching Effectiveness: Cognitive and Affective Criteria." *Journal of Educational Psychology* 72: 468-75.
Marsh, Herbert, J. U. Overall, and S. P. Kesler. 1979. "Validity of Student Evaluations of Instructional Effectiveness: A Comparison of Faculty Self-Evaluations and Evaluations by Their Students." *Journal of Educational Psychology* 71: 149-60.
Miller, Richard I. 1987. *Evaluating Faculty for Promotion and Tenure.* San Francisco: Jossey-Bass.
Morrisey, George L. 1983. *Performance Appraisals in the Public Sector: Key to Effective Supervision.* Reading, MA: Addison-Wesley.
Morsh, J. E., G. C. Burgess, and P. N. Smith. 1956. "Student Achievement as a Measure of Instructional Effectiveness." *Journal of Educational Psychology* 47: 79-88.
Mosher, Fredrick C. 1982. *Democracy and the Public Service,* 2nd ed. New York: Oxford University Press.
Norr, J. L., and K. S. Crittenden. 1975. "Evaluating College Teaching as Leadership." *Higher Education* 4: 335-50.
O'Sullivan, Elizabethann, and Gary R. Rassel. 1989. *Research Methods for Public Administrators.* New York and London: Longman.
Parten, M. 1950. *Surveys, Polls, and Samples.* New York: Harper & Row.
Rich, Wilber C. 1989. "Appraising Employee Performance." In *Handbook of Public Administration,* ed. James L. Perry. San Francisco and London: Jossey-Bass.
Rutland, Peter. 1990. "Some Considerations Regarding Teaching Evaluations." *The Political Science Teacher* 3: 1-2.
Seldin, Peter. 1984. *Changing Practices in Faculty Evaluation: A Critical Assessment and Recommendations for Improvement.* San Francisco: Jossey-Bass.
Student Evaluation of Teaching Advisory Committee. 1986. "A Proposal for a Student Evaluation of Teaching Form for Personnel Decisions," unpublished report, The Ohio State University.
The University of Toledo. 1987. "Report of the Committee on Teaching Evaluation of the College of Arts and Sciences."
Webb, W. B., and C. Y. Nolan. 1955. "Student, Supervisor, and Self-Ratings of Instructional Proficiency." *Journal of Educational Psychology* 46: 42-46.
Welch, Susan, and John Comer. 1988. *Quantitative Methods for Public Administration,* 2nd ed. Chicago: The Dorsey Press.
Wiggins, J. S. 1973. *Personality and Prediction: Principles of Personality Assessment.* Reading, MA: Addison-Wesley.
Wilson, O. W. 1972. *Police Administration.* New York: McGraw-Hill.

About the Author

Hugh Hinton is an associate professor of political science and public administration and former director of the public administration program at The University of Toledo.