Woman C.P.A.

Volume 41 | Issue 3 Article 8

7-1979

Education: Faculty Evaluations - Value and Validity

Carole Cheatham

Education

Faculty Evaluations

Value and Validity

Editor:
Carole Cheatham, CPA, Ph.D.
Mississippi State University
Mississippi State, Mississippi

For many years it has been the task of college teachers to evaluate students and to assign them a grade of A, B, C, D or F according to their performance. In more recent times, particularly since the student unrest of the midsixties, the students have in turn evaluated the teachers at many institutions, frequently also on a five-point scale and have assigned them grades of five to one. This evaluation by students is usually done on a standardized form. Some methods use class time toward the end of the semester, some require that the forms be mailed in toward the end of the semester or after the semester is over, and some require the evaluation be done in the early days of the following semester.

Typical of the items on which students rank their instructors are: "The instructor's objectives for the course have been made clear," "The instructor used class time well," "The instructor was readily available for consultation with students," "Lectures were too repetitive of what was in the textbook," "The instructor was enthusiastic when presenting course material," "The text was clear in presentation of concepts."

Some schools have designed their own forms and some have elected to use a standardized form and rating scale such as the Educational Testing Service form based on the Michigan State University scale, the form from the Berkeley Center for Research and Development on Higher Education, the Purdue Rating Scale for Instruction, or the Illinois Course Evaluation Questionnaire. The standardized forms have the advantage of being more thoroughly researched and of allowing comparability with other institutions. The self-developed forms are more adaptable to a particular situation and may be less expensive because they need not be purchased from an outside source.

Purposes of Faculty Evaluations

Basically there are three purposes of faculty evaluations: 1) to help faculty members improve their instruction techniques, 2) to guide students in their selection of courses and/or teachers, and 3) to assist administrators in their evaluation of the teaching abilities of individual instructors. To these purposes may be added a somewhat auxiliary purpose: 4) to conduct research on faculty performance.

The first purpose, that of assisting instructors in self-improvement, is certainly a worthwhile goal and was probably the first motivation toward faculty evaluations. Long before the days of standardized and compulsive evaluations, some teachers were designing and administering their own questionnaire in an honest attempt at improvement. Provided these faculty members were not so blind to their shortcomings that they failed to ask the right questions, they received meaningful information that assisted them in bettering their instruction techniques. However, the teachers that needed the most improvement were usually those that failed to ask for or ignored any kind of feedback from their students. Consequently, faculty evaluations were only made of some conscientious teachers who were motivated to improve, and they were probably good teachers anyway. In order to get the message across to the poorer teachers it was necessary to make the evaulations compulsory and to provide some sort of standardized form for general use. On the whole, the poorer teachers ignored the results from these evaluations as they ignored less formal forms of feedback.

The second purpose of faculty evaluations is to guide students in course and/or faculty selection. Using a faculty evaluation for this purpose formalizes a process in which students have always engaged and provides an information supply with equal access for all students. While formerly students had to rely on word-of-mouth or the informal files of a sorority or fraternity, they could now consult a handbook or their college library to obtain this information. This assumes, of course, that the results of evaluations were made available to students, which is not the case at all institutions.

The major objection to using faculty evaluations to guide students in their course and faculty selection relates to the confidentiality of the information. Some faculty members are sensitive about having their ratings generally known. Those who object to publishing faculty ratings point out that the confidentiality of student grades is protected by the Buckley Amendment and should not instructors have the same rights to privacy?

The third purpose of faculty evaluations is to assist administrators in their evaluation of teaching ability. This is probably the most controversial use of faculty evaluations. One has to sympathize with an administrator who must make decisions concerning promotion, tenure, and salaries given the

information at his/her disposal. Teaching, which is *the* major or at least *a* major activity of faculty, is not easily assessed. Self-evaluations by instructors have obvious difficulties. Classroom visits tend to provide very poor samples of performance besides being grossly unpopular. Achievement tests tend to apply only in courses stressing rote learning. Peer ratings sound good, but as far as classroom performance is concerned they can only be based on hearsay — which is what the administrator would probably base his/her evaulation on anyway. Given the alternatives, student evaluations of faculty seem the ideal answer.

Why then the strenuous objections by some faculty members to this method? Most of the objections center around validity. One proponent of teacher evaluations by students quotes from Aristotle's *Politics* which declares that we receive a better notion of the dinner from the guests than from the cook, likening the students to the guests and the teacher to the cook.[1] This may be true. However, the guests are far more likely to give an opinion based on flavor than on nutrition; and, in the long run, it is nutrition that counts. Opponents to this use of faculty evaluations say that students tend to give an opinion of a course or an instructor based on how much they enjoyed it rather than on what they learned from it. The charge is that faculty evaluations measure popularity rather than teaching ability.

The fourth purpose of faculty evaluations is to conduct research on factors related to faculty performance. This was listed above as an auxiliary purpose because most of the research done with faculty evaluations is to prove or disprove the validity of the instrument rather than assess performance. In other words, the research has been the result rather than the cause.

Validity of Faculty Evaluations

The most serious charge against faculty evaluation instruments is that they lack validity. That is, that they do not measure what they purport to measure — teaching effectiveness.

There are many factors that influence the rankings given by students in faculty evaluations. Where these variations are known and allowed for in the interpretation of the results, the rankings are still usable. Some of these factors relate to the questionnaire itself and the way it is administered. Students may react negatively to an overly long questionnaire. They are most likely to complete a questionnaire with clear instructions and easy to check answers.[2] Students may also react to teaching conditions over which the teacher has little control. In general, research has shown that lower-level courses, moderate-sized classes, and required courses tend to receive less favorable ratings.[3] Classes held during the middle of the day receive higher rankings than those held in the early morning.[4]

Dr. Fox and Other Interesting People

There are other factors that influence rankings that are more subtle and harder to allow for in the interpretation of rankings. Consequently, these defects are of a more serious nature. One assertion is that faculty evaluation results are unduly influenced by the "popularity" of the instructor. There are several studies that appear to confirm this.

Williams and Ware conducted a study in which they hired a Hollywood actor to deliver six types of lectures. The content density was high, medium or low. The manner of delivery was high expressive or low expressive. In the high-expressive lectures the actor used devices such as humor, enthusiasm, and voice modulation while not using them in the low-expressive lectures. Afterwards students were administered an achievement test and asked to rank the lecturer. As might be expected, high scores on the achievement test were associated with high content. High rankings of the lecturer were associated with high expressiveness. In Williams and Ware's early study in 1975, it appeared that high expressiveness also aided achievement, but this was not born out in a later study. The correspondence between high expression and high rankings of the instructor without regard for content is what the authors termed a "Dr. Fox effect."[5]

Keaveny and McGann (1978) did a study relating student ratings to certain behavioral clusters. Two clusters related to competence and organization which the authors labeled "Taut Ship" for high levels and "Loose Ship" for low levels. Another two clusters related to concern and consideration which the authors labeled "Nice Guy" for high levels and "Bad Guy" for low levels. As might be expected, "Nice Guy-Taut Ship" received the highest overall ratings and "Bad Guy-Loose Ship" received the lowest overall ratings. However, "Nice Guy-Loose Ship" had a better chance of a good overall rating than did "Bad Guy-Taut Ship", indicating that students appear to be more influenced by the consideration variables than the competence variables.[6]

Since certain variables appear to affect the outcome of faculty evaluations, one author has suggested that an instructor might use these effect to "cheat" on the evaluation. In a rather tongue-in-cheek article Michael Faia suggests:

> *As in the case of student cheating, the more interesting techniques are the more subtle ones. To begin with, we must make use of the findings of social psychology. For instance, research shows that course evaluations are influenced by a host of factors that have nothing to do with the "objective" aspect of teaching, such as whether or not professors are married, how they dress, whether they act "seductively" (as in the famous "Professor Fox" experiments), whether or not professors share the values of their students, whether or not students receive the grades they expect, whether or not instructors show "hostility."*[7]

Grade-Rankings Correlation

Besides the assertion that rankings are influenced by a group of behavioral variables that may loosely be characterized as "popularity", there is also the assertion that rankings are unduly influenced by the grade that a student receives or expects to receive in a course. This claim crops up over and over with good reason. A correlation between rankings and grades has occurred in many major studies.

Table 1 presents the findings from twenty-nine large grade-rating studies published between 1934 and 1974. Twenty-eight of the studies show positive correlations between grades given students and rankings given instructors. In total the studies represent more than 80,000 student ratings in thirty-five or more colleges and universities. The only study of this group which shows a negative correlation is the Heilman and Armentrout study which was done in 1935, and it is open to serious question from a control standpoint because the teachers apparently administered and handed in their own rankings.[8]

Table 1 does not present an exhaustive list of all the studies that have

TABLE 1
PUBLISHED DATA FROM 29 LARGE GRADE-RATINGS STUDIES
1934 — 1974

Author and Date of Publication	Maximum Grade-Rating Correlation Found
1. Anikeef (1953)	+ coefficient of .73 in freshman-sophomore classes
2. Bassin (1974)	+ coefficient of .10 affecting rankings to 32 percentiles
3. Bausell & Magoon (1972)	+ coefficient of .6
4. Centra & Linn (1973)	+ correlation; unstated "moderate" amount
5. Cornwell (1974)	+ correlation accounting for 11% of variance
6. Echandia (1964)	+ correlation at .01 level of significance; no coefficient given
7. Elliott (1950)	+ correlation on all 10 items on Purdue rating scale; no coefficient given
8. Granzin & Painter (1973)	+ coefficients of .14 to .21
9. Heilman & Armentrout (1936)	—coefficient of .04
10. Hildebrand, et al. (1971)	+ coefficient; unstated amount
11. Holmes (1971)	+ correlations: 5 to 11% of variance
12. Hudelson (1951)	+ coefficient of .19
13. Kennedy (1972)	+ correlation significant at .01 level; no coefficient given
14. Kooker (1968)	+ correlation at .001 level; no coefficient given
15. Mirus (1973)	+ coefficient of .85
16. Nichols & Soper (1972)	+ coefficient of .53
17. Overturf & Price (1966)	+ coefficient of .17; questionable statistical method used
18. Perry & Baumann (1973)	+ correlation of .78
19. Rayder (1968)	+ coefficient of .18
20. Rosenshine, et al. (1973)	+ correlations of .09 to .27
21. Rubenstein & Mitchell (1970)	+ correlations of .09 to .44
22. Spencer & Dick (1965)	+ coefficient of .85 to .91 in one study; + correlation of unstated amount in second study
23. Starrack (1934)	+ coefficient of .15
24. Stewart & Malpass (1966)	+ correlation significant at .001 level; no coefficient given
25. Voeks & French (1960)	+ coefficients up to .60 in one study; + correlations in 9 of 10 departments in second study; indeterminant results in third study because of faulty design
26. Walker (1969)	+ coefficient of .48 by rank order
27. Weaver (1960)	+ correlation significant at .001 level; no coefficient given
28. Powell (1974)	+ coefficient of .73
29. Powell (1975)	+ coefficient of .79

Source: Robert Powell, *College English,* January 1978, pp. 628-629.

been done in the area of faculty evaluations, and there are studies that demonstrate negative or no correlation between rankings and grades. However, some of these studies were done by evaluation consultation services which have a vested interest in proving the validity of their tests. Some other studies involve situations in which the teacher did not control the students' grades. Some negative correlation studies or no correlation studies were very small involving as few as one teacher. (This can also be said of some studies which found positive correlation although all those that appear in Table 1 involve at least five teachers.)[9]

Attempts to Establish Validity

The claim of lack of validity is indeed a serious claim and this claim has not been adequately refuted by the proponents of teacher evaluations. Attempts to deal with the problem have taken several forms. Consider, for example, the statement from a book published by one firm specializing in

evaluation programs, which presents three methods for testing validity:

> *The validity of an instrument, or whether it measures what it purports to measure, has been studied extensively for some instruments. Other institutions pilot test their own instruments, and may test the validity by requesting the same information in a variety of ways on different items, and then seeing if the answers are statistically consistent ... Validity is often measured by comparing a test intrument with one that has already established its validity. Many committees decide that face validity is acceptable; that is, the instrument logically appears to be valid.*[10]

The first method of testing for validity, that of asking for the same information in a variety of ways, is certainly a useful way to establish validity although its use with a single instrument is limited due to considerations of length. However, as regards the second method, testing an intrument with another valid instrument is not possible until it is established that there is a valid standard for teacher evaluations. Accepting a questionnaire on the basis of face validity, the third method, is like an auditor giving a clean opinion of a balance sheet because the figures add up. Equally unimpressive are items on the survey form such as "I have given thoughtful consideration to the questions on this form,"[11] which only prove the student read the item.

Conclusions

Teacher evaluations have been used for four purposes — for teacher self-improvement, for student guidance in selecting teachers and/or courses, for assessment of teachers' performance by administrators, and for research purposes. It appears that teacher evaluations do have some use for teacher self-evaluation particularly in regard to single items asked on the forms. For example, if a teacher consistently gets low rankings on an item such as "Spoke with expressiveness," he or she can strive for improvement in that area. Interpretations of overall rankings should be tempered by the knowledge that variables other than teaching effectiveness do affect these rankings.

Use of faculty evaluations by students to select courses is a valid use although permission of the instructor should be obtained in order to respect the confidental nature of the rankings. For the typical student seeking a professor and/or course the rankings are probably fairly accurate, assuming his or her goals and reactions will be similar to those of previous students. For the student with atypical goals and reactions, the rankings will be less useful.

Use of faculty evaluations by administrators is probably unwise in view of the lack of established validity. It is particularly hazardous to compare one faculty member's rankings with those of another faculty member. If it is desired to assess teaching effectiveness, then achievement tests administered to students appear to be more to the point, although achievement tests have problems also. Perhaps the only feasible alternative at present is to continue to rely largely on more objective measures of performance such as publications, offices held, committees chaired, etc. If and when more valid teacher evaluation instruments are developed, then they can be utilized. Re-testing the present survey forms appears to be of limited value because most have been tested extensively, and their validity is still in question. More research needs to be done to develop better measures of teaching effectiveness, perhaps utilizing achievement tests or some combination of achievement tests and student rankings. □

NOTES

[1]Richard I. Miller, *Developing Programs for Faculty Evaluation*. San Francisco: Josey-Bass Publishers, 1974, p. 30.

[2]William J. Genova et al., *Mutual Benefit Evaluation of Faculty and Administrators in Higher Education*. Cambridge, Mass.: Ballinger Publishing Co., 1976, p. 29.

[3]Ibid.

[4]Miller, p. 66.

[5]Reed G. Williams and John E. Ware, Jr., "Validity of Student Ratings of Instruction Under Different Incentive Conditions: A Further Study of the Dr. Fox Effect, "*Journal of Educational Psychology*, Vol. 68, No. 1, pp. 48-56.

[6]Timothy J. Keaveny and Anthony F. McGann, "Behavioral Dimensions Associated with Students' Global Ratings of College Professors," *Research in Higher Education*, December 1978, Vol. 9, pp. 333-345.

[7]Michael A. Faia, "How — and Why — to Cheat on Student Course Evaluations," *Liberal Education*, March 1976, Vol. LXII, No. 1, p. 118.

[8]Robert Powell, "Faculty Rating Scale Validity: The Selling of the Myth," *College English*, January 1978, Vol. 39, No. 5, pp. 616-619.

[9]Kenneth Elbe and Robert Powell, "Comment and Response," *College English*, January 1979, Vol. 40, No. 5, p. 564.

[10]Genova, p. 29.

[11]Student Reaction to Instruction and Courses (IDEA), Center for Faculty Evaluation and Development in Higher Education, Manhattan, Kansas, 1975.