The Reliability of Students' Ratings of Faculty Teaching Effectiveness.

Published in: College Teaching, Summer2001,Professional Development Collection

Section:

COMMENTARY

Procedures for measuring faculty teaching effectiveness vary by university; however, student evaluations typically are considered in the process and are critical elements of tenure and promotion decisions (Haskell 1997; Marsh 1987). Because of the high correlation between quality teaching and high student achievement (Darling-Hammond 1997), it is understandable that faculty teaching effectiveness be carefully monitored. Student evaluations of faculty teaching effectiveness are also used in dispensing merit-based salary increases and can create a competitive climate among faculty members within university colleges and departments. Because of the emphasis placed on student evaluations and the pressure for junior faculty members in particular to receive high ratings to bolster their promotion and tenure documents, an examination of the reliability of student evaluations of teaching effectiveness is warranted.

The weight that the student evaluations receive differs across universities and is continually under scrutiny (Haskell 1997; Sproule 2000). Faculty concerns regarding the use of student-completed evaluation forms as the sole or most important assessment of teaching quality have been well documented (Cashin 1983; Haskell 1997; Mark 1982; Marsh 1987). Most faculty agree on the need for student input into the evaluation of their teaching; however, some express concern over student ratings as the sole evaluation criterion and question the reliability and validity of these measures. Mixed results from tests of the reliability and validity of student evaluations of faculty underscore that concern (Haskell 1997). Simmons (1996) reported two separate studies, both of which found poor reliability for student-completed evaluation instruments. Instrument problems included ambiguous items, positively or negatively skewed items, and items that had no correlation to classroom teaching performance. Earlier research on the validity and reliability of student-completed course evaluation instruments reported the instruments to be reliable, stable, and relatively valid (Marsh 1987; Marsh and Bailey 1993; Peterson and Kauchak 1982). However, a review of

the research revealed reliability and stability over long periods of time and in multiple courses for the same instructor. None of the research examined the reliability of the individual students.

A review of the research uncovered two documented measures of the reliability of student evaluations. First, Marsh (1987) found average student evaluations to be reliable and stable among students within a specific class and determined reliability of instructor ratings by examining ratings from different courses taught by the same faculty member. Second, students evaluated the instructor and the learning environment that he or she created, not the course, thus lending credence to the long-term stability of student evaluations of faculty teaching established by the aggregate scores of all students in a class over several semesters (Marsh and Bailey 1993; Peterson and Kauchak 1982).

The study that we report in this article differs from previous research in two ways. First, instead of looking at an instructor's ratings over time, comparing class-average student ratings, or examining the instrument itself, we explored the reliability of the student as an evaluator. Second, we used teacher education students as the sample population. While simultaneously earning a baccalaureate degree and professional licensure, teacher education students enroll in an interesting mix of courses, ranging from liberal arts classes, such as "Introduction to Philosophy," to teacher education courses, such as "Methods for Teaching Secondary Social Studies."

The research questions guiding the study were:

1. Do individual students reliably evaluate the effectiveness of the courses in which they are enrolled? Specifically, do student ratings of teaching effectiveness match their holistic impression of the quality of the course?
2. Is there a difference between student evaluations of teacher education courses and student evaluations of non-teacher education courses?

Method

This study used three different measures of teaching effectiveness to examine the reliability of student evaluators. In this study, teacher education students composed the sample population, evaluating both teacher education and non-teacher education courses.(n1)

Setting and Participants

The study took place at a midsized Western university. We invited 202 students either enrolled in or interested in a teacher education program [the teacher education program is housed in the second-largest college at the university] (elementary, elementary/special, and secondary education) to participate (see table 1), and 128 agreed for a return rate of 63 percent. Representative of the gender ratio in most teacher education programs, substantially more females (ninety-four) than males (twenty-seven) participated in the study. Eight participants did not indicate gender. Forty participants were in a secondary teacher education program, and eighty-six participants indicated that they were in an elementary education, special education, or combination program. Students in the special education and the combination of special and elementary education were reported together. Two participants did not indicate any teacher program or university major.

Data Collection

In the last two weeks of a semester, we approached faculty members teaching introductory or general methods courses for the elementary, special education, and secondary programs, and all of them granted us permission to invite students to participate in the study. We attended five of eight sections of five different introductory education courses. We chose introductory courses specifically because the students enrolled in these lower-division course typically were also enrolled in a number of courses outside of teacher education. Reading from a prepared and Institutional Review Board-approved script, we explained the project and instructions for completing the instruments. After we left the classroom, students were given approximately thirty minutes to complete the instruments. To ensure the confidentiality of the participants, all students were required to return the instruments, whether or not they had completed the forms.

Instrumentation

Students were asked to complete three measures of teaching effectiveness. First, students were asked to rank all of the courses they were enrolled in that semester holistically, from "most effectively taught" to "least effectively taught." We did not define an "effectively taught course." Second, students completed an attribute rating scale for each course that they had rated holistically (see table 2). The attribute rating scale consisted of twenty items on a five-

point, Likert-type scale (5 representing most effective). The attribute rating scale had been used for the previous three years as the standard course evaluation form in all three departments of the College of Education. The results of these course evaluations are used by the college and university administration as the primary indicator of a faculty member's teaching effectiveness for annual reviews (merit is based on scholarship, teaching, and service) and for promotion and tenure decisions. No university-wide instrument was available for use. The number of attribute rating scales completed varied among students (M = 4.38; SD = 1.19), with students completing between one and six attribute rating scales. As a result, 128 participating students completed 417 attribute rating scales. The third measure of teaching effectiveness was a writing prompt. For each course evaluated, students were asked to provide a narrative reply to the prompt, "How would you describe this course to a friend?" This prompt resulted in 417 narrative responses.

Findings

We calculated a mean score for each twenty-item attribute rating scale completed (n = 417). Mean scores for each student then were ranked and matched with his or her holistic scores. This created a data set for each student that aligned each course's holistic rank with its corresponding mean score on the attribute rating scale (n = 111). Students who listed only one course on the holistic rank or completed only one attribute rating scale were not part of the analysis. Initially, the data set also contained the narrative comments. However, due to a large number of unrelated and/or instructor-identifying comments, the narrative comments were not part of further analysis. Therefore, the holistic rank scores and the mean scores of the attribute rating scales were the data used for analyses.

To answer the first research question regarding student reliability in rating courses, we compared the two aligned data sources (holistic ranks and mean scores on the attribute rating scale) for each course and created a new data file for each student. We displayed the student's holistic ranking of courses from "most effectively taught" to "least effectively taught" (see table 3). The mean scores on the attribute rating scale then were displayed from highest mean score to lowest mean score. The two displays for each student were compared to one another to determine the degree to which they matched. In other words, this comparison determined how often the course ranked "most effectively taught" also had the highest attribute rating scale mean score. This match was converted to a percent and added

to the data set for each student. If the course ranked "most effectively taught" had the highest mean score, the next "most effectively taught" had the second highest mean score, and the remaining two courses also matched, the agreement was 100 percent. When the numbers for the two data sources did not match, the agreement was less than 100 percent. In terms of the overall sample, the correlation between the attribute rating scale and the holistic rank was significant (Kendall Tau Coefficient = .43; $p < .05$) and supports previous research (Marsh 1987). However, this relationship does not reveal the reliability of student evaluators on an individual level.

In examining the results of the two data sources for individual students, we found that the matches broke naturally into thirds. Analysis of each individual student's data set revealed that 29 percent of the students matched the holistic ranks to the attribute scale for 100 percent of their courses. Eight percent of the students' attribute scores matched their holistic ranks for 67 percent to 99 percent of their courses. A 67 percent match means that a student rank ordered six courses and completed six attribute rating scales, with four of the six rank-ordered positions matching. For 32 percent of the students, the match between data sources occurred in less than 34 percent of their courses. Finally, 11 percent of the students did not match a holistic rank with an attribute mean score for any course evaluated.

The second research question compared teacher education and non-teacher education attribute rating scales completed by students. Teacher education courses were defined as those in the teacher education division of the College of Education. These included introductory courses, methods courses, and multicultural education courses. Non-teacher education courses were all other courses in which students were enrolled. These included courses in the arts and sciences as well as other courses in the College of Education that were not in the teacher education division (e.g., educational psychology and school law). The mean evaluation scores for the teacher education rated courses ($M = 3.950$; $SD = .815$) compared to the mean evaluation scores for non-teacher education rated courses ($M = 3.695$; $SD = .806$) revealed a statistically significant difference ($t\ 478 = 3.44$; $p = <.05$). Overall, students evaluated teacher education courses higher than non-teacher education courses on the attribute rating scale.

Discussion

The initial analysis of the first research question supported the previous findings endorsing the reliability of student evaluations of faculty teaching. However, on further analysis at the individual student level, those results were of concern, as it appeared that what students holistically ranked as the "most effectively taught" course did not correspond with the results on the attribute rating scale. The assumption was that a reliable student evaluator was one whose overall ranking of the most effectively taught course matched the highest mean score on the attribute rating scale, so that the course ranked least effectively taught would have the lowest mean score. We found this to be true for less than one-third of the students sampled. Whereas previous studies looked for reliability across a group of students and over time, the comparison in this study examined the reliability of the individual student. However, if, as this study found, individuals are not consistent in their evaluations, then aggregated reliability measures are giving faculty a false sense of security. The integrity of the student as evaluator in the tenure and promotion process should be studied more closely.

On closer examination of the data, we began to question the validity of the instruments used to measure faculty teaching effectiveness. Specifically, students' holistic rankings represented their own perceptions of quality teaching with no parameters set by a standardized evaluation instrument. Kishor (1995) contends that students base their evaluations on an implicit personality theory of a good instructor. When given no specific parameters (i.e., the attribute rating scale), students recall previous information and infer other information to form their personality theory about good instructors. This theory then is applied to each instructor, searching for a goodness of fit. It is a reasonable assumption that participants in this study used their own implicit personality theories in ranking their courses holistically. A mismatch occurred between the characteristics that an individual student perceived to be indicative of an effectively taught course and the characteristics that the university, college, or department asked him or her to consider. Universities currently use student answers to questions about traits that the institution values. This practice forces students to consider teaching effectiveness through the lens of the institution rather than through their own personal lens. Attribute rating scales are intended to give students a voice. However, that voice may be inauthentic.

The reliability of student evaluators may be further confounded by research indicating that student-completed evaluations measure "popularity of the instructor" rather than "teaching effectiveness" (Altschuler 1999; Sproule 2000). Teachers perceived as enthusiastic, good-

humored, and warm consistently fare better on student evaluations. Although these characteristics are pleasant, they do not equate with teaching effectiveness. As stated earlier, student-completed evaluations are more about the instructor than about the actual course (Marsh 1987). The results, then, could be not just an issue of unreliable student evaluators or an invalid instrument. Rather, the system for evaluation itself may be inconsistent. This system requires that students use an instrument corresponding with the institution's definition of teaching effectiveness, rather than the students' definitions of teaching effectiveness. This inconsistency only becomes evident when students complete multiple measures, including measures that reflect the institution's definition and ones that reflect the students' definitions.

The findings related to the first research question temper the discussion of the second. If individual students' evaluations are not reliable from instrument to instrument, as found in this study, significance of the results related to the second question may not be meaningful. We found a significant difference between teacher education students' evaluations of teacher education courses and non-teacher education courses. Using the twenty-item attribute rating scale, teacher education students rated teacher education faculty higher than non-teacher education faculty. There are multiple possible explanations for this difference. One possibility is that teacher education faculty tend to have more training and experience in creating a positive learning environment. Teacher education faculty have made a career of researching and modeling effective teaching practices. A second and related possibility is that teacher education faculty may value teaching, particularly to undergraduates (the teacher education courses in this study), more than non-teacher education faculty. If one defines teacher education as a soft or an applied discipline, then the results of this study concur with previous research reporting a fairly consistent pattern of student-completed evaluations that favored soft and applied disciplines over math, engineering, and many sciences (Cashin 1990; Feldman 1978). Smart and Elton (1982) provide a rationale for the difference; they found that faculty in soft and applied disciplines place a greater emphasis on the process of teaching. A third possibility is that teacher education classes are desirable to these particular students. They are majoring in education or at least exploring the field; therefore, interest is high.

Other less-likely possibilities are related to the course (e.g., class size, required or elective, lower or upper division) and/or student characteristics (e.g., gender, class rank, ability). Although assumptions may be that students evaluate more positively courses in their major

and smaller classes, existing research indicates that the relationships among course characteristics, student characteristics, and faculty evaluations are mixed at best. For example, Centra (1981) found that the relationships among course and student characteristics and student evaluations are very small and generally insignificant. In addition, Marsh (1987) and Simmons (1996) reported that smaller classes tend to be evaluated more highly and less reliably, although Feldman (1978) found no significant difference by class size.

Conclusion

Previous research found that student evaluations of faculty teaching are reliable measures (Marsh 1987; Marsh and Bailey 1993; Seldin 1984); however, reliability was established by several groups of students rating a faculty member consistently over a period of time, much like test-retest reliability. What remains unclear is the reliability of individual students in evaluating faculty teaching effectiveness, whether the reason is the unreliability of the student, an invalid instrument, or an inconsistent system. Therefore, the purpose of our study was to examine the ability of individual university students to evaluate courses consistently and to rate courses on a teaching effectiveness rating scale in accordance with their own perceptions of the most and least effectively taught courses. In addition, the difference between student evaluations of teacher education and non-teacher education courses was of interest. Our findings are mixed. Analysis of the overall sample indicated student reliability between two different measures of teaching effectiveness, but analysis of each individual student's reliability raised concern.

There are many ways to make evaluation instruments, as well as overall evaluation of faculty teaching effectiveness, more valid and reliable. Similar to Feldman's conclusions (1988), the results of this study indicate that what students perceive as effective teaching may not correspond with what the institution perceives as effective teaching. If institutions continue to believe in the importance of the student voice in evaluating faculty, it may be necessary to include students in the construction of the evaluation instrument. It also may be important to determine exactly what students are evaluating and whether they are able to evaluate teaching effectiveness based on their experience and expertise. That is, are students evaluating teaching effectiveness or something more akin to the likability of the instructor, course content, or topic? Another solution is the use of multiple measures (e.g., peer evaluation, administrator evaluation, self-evaluation), which is recommended by many

researchers who contin,e to support the use of student evaluations as one measure of faculty teaching effectiveness (Cashin 1983; Haskell 1997; Mark 1982; Marsh 1987).

The issue of the reliability of student evaluators of university faculty warrants further examination. Research should be extended to other universities and other populations (beyond teacher education students). The results of this study may be of interest to faculty whose teaching is evaluated solely by students. Considering the potential implications for faculty tenure, promotion, and financial merit decisions, the wholesale support of student evaluations of faculty teaching invites debate and continued consideration.

NOTE

(n1.) For those interested in a more detailed description of the method and results, please contact the first author, Kathryn Obenchain, at kmo@unr.edu.

Table 1.--Number of Students Enrolled in Teacher Education Program

Legend for Chart: B - Elementary C - Elementary/Special education D - Secondary E - Total(a)

A	B	C	D	E
Gender Male	7	4	15	27
Female	41	30	22	94
Age 18-22	31	19	15	66
23-27	9	4	10	23
28-32	3	2	7	12
33-37	4	5	4	13
38+	4	4	4	13

(a) Program totals that do not sum to the overall total reflect missing data.

Table 2.-Attribute Rating Scale Items

1. Material presented was challenging and stimulating.

- 2. I was interested in the subject.
- 3. The instructor displayed enthusiasm in teaching the subject.
- 4. I felt motivated to do my best work.
- 5. Examples and illustrations were used effectively.
- 6. The instructor explained what is expected of students.
- 7. A clear course outline or syllabus was provided.
- 8. The instructor clarified material that needed explanation.

- 9. The subject was presented effectively.
- 10. The instructor referred to recent applications or developments in this area of study.
- 11. The instructor related to students in the class in a way that prompted mutual respect.
- 12. The instructor was available outside the class to individual students for help and/or advice.
- 13. Evaluation procedures were consistent with the syllabus.
- 14. The course had relevant assignments and other requirements.
- 15. The instructor gave helpful feedback on my performance.
- 16. In-class learning activities and discussion engaged students with subject matter.
- 17. Outside-of-class assignments contributed to my understanding of the subject matter.
- 18. Effective textbooks and other materials were used.
- 19. The knowledge gained in this course will contribute to my present or future career.
- 20. Overall, this course was worthwhile.

Table 3.--Sample Student Holistic Course Ranking and Attribute Rating Scale Comparison

Legend for Chart: A - Holistic rank B - Attribute rating scale mean score C - Match(a) A B C 1 4.80 Match 2 4.20 No match 3 4.30 No match 4 3.75 Match 5 2.85 Match 6 2.75

Match (a) The holistic rank matched the attribute rating scale score for 67 percent (four of six) of the student's courses.

Altschuler, G. 1999. Let me edutain you. The New York Times, Education Life Supplement, April 4.

Brown, R. 1977. The relationships between student evaluation of teaching, student achievement and student perception. ERIC Document Reproduction No. ED 133 314.

Cashin, W. E. 1983. Concerns about using student ratings in community colleges. New Directions for Community Colleges 11(1): 57-65.

-----. 1990. Students do rate different academic fields differently. In Student ratings of instruction: Issues for improving practice, ed. M. Theall and J. Franklin, 113-22. San Francisco: Jossey-Bass.

Centra, J. A. 1981. Determining faculty effectiveness. San Francisco: Jossey-Bass.

Darling-Hammond, L. 1997. Doing what matters most: Investing in quality teaching. New York: National Commission on Teaching and America's Future.

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

-----. 1988. Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? Research in Higher Education 28:291-344.

Haskell, R. E. 1997. Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. Education Policy Analysis Archives 5(6).

Kishor, N. 1995. The effect of implicit theories on raters' inference in performance judgment: Consequences for the validity of student ratings of instruction. Research in Higher Education 36(2): 177-95.

Mark, S. F. 1982. Faculty evaluation in community college. Community Junior College Research Quarterly 6(2): 167-78.

Marsh, H. W. 1987. Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. [Monograph summary]. International Journal of Educational Research 11:253-387.

Marsh, H. W., and M. Bailey. 1993. Multidimensional students' evaluations of teaching effectiveness: A profile analysis. Journal of Higher Education 64(1): 1-18.

Peterson, K., and D. Kauchak. 1982. Teacher evaluation: Perspectives, practices, and promises. Salt Lake City, Utah: Utah University Center for Educational Practice.

Seldin, P. 1984. Changing practices in faculty evaluation. San Francisco: Jossey-Bass.

Simmons, T. L. 1996. Student evaluation of teachers: Professional practice or punitive policy? JALT Testing & Evaluation N-SIG Newsletter 1(1): 12-6.

Smart, J. C., and C. F. Elton. 1982. Validation of the Biglan model. Research in Higher Education 17(3): 213-29.

Sproule, R. 2000. Student evaluation of teaching: A methodological critique of conventional practices. Education Policy Analysis Archives 8(50).

~~~~~~~~

By Kathryn M. Obenchain; Tammy V. Abernathy and Lynda R. Wiest

Kathryn M. Obenchain and Tammy V. Abernathy are assistant professors, and Lynda R. Wiest is an associate professor in the Department of Curriculum and Instruction at the University of Nevada in Reno.
~~~~~~~~