strain and change along these lines. The inability of men in the chain of command to appreciate the charges brought by women who claimed to have been sexually molested at a Tailhook convention led to the early retirement, resignation, and sanctioning of several high-ranking officials in the Department of the Navy. Now women are being trained as combat pilots in the Air Force, the first branch of the military ever to be led by a female service secretary.

12. I use overhead projections rather than handouts so that I can control the presentation. With handouts, students tend to skip ahead so that the whole class is not looking at the same illustration simultaneously, and those who are having trouble seeing an illustration in more than one way can feel left behind.

References

Bloomer, Carolyn M. 1976. *Principles of Visual Perception*. New York: Van Nostrand Reinhold.

Cassel, Jeris F., and Robert J. Congleton. 1993. *Critical Thinking: An Annotated Bibliography*. Metuchen, NJ: Scarecrow Press.

Cohen, Mel. 1993. "Making Critical Thinking a Classroom Reality." *PS* 26.

Coren, Stanley, and Joan Stern Girgus. 1978. *Seeing Is Deceiving: The Psychology of Visual Illusions*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ennis, Edward. 1962. "A Concept of Critical Thinking." *Harvard Educational Review* 32.

Freidman, Thomas L. 1990. "A Dreamlike Landscape, a Dreamlike Reality." *New York Times*, 28 October.

Glaser, Edward M. 1941. *Experiments in the Development of Critical Thinking*. New York: AMS Press.

Greenfield, Patricia, and Paul Kibbey. 1993. "Picture Imperfect." *New York Times*, 1 April.

Hatsumi, Reiko. 1993. "A Simple 'Hai' Won't Do." *New York Times*, 15 April.

Lewis, Flora. 1990. "Between-Lines Disaster." *New York Times*, 19 September.

Luckiesh, M. 1965. *Visual Illusions: Their Causes, Characteristics, and Applications*. New York: Dover.

Paul, Richard W. 1992. *Critical Thinking: What Every Person Needs to Survive in a Rapidly Changing World*, ed. A.J.A. Binker. Santa Rosa, CA: Foundation for Critical Thinking.

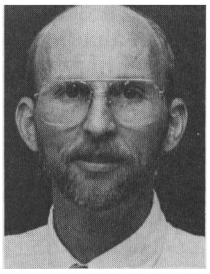
Rosenthal, Andrew. 1990. "Did U.S. Overtures Give Wrong Idea to Hussein?" *New York Times*, 19 September.

Tannen, Deborah. 1991. *You Just Don't Understand: Women and Men in Conversation*. New York: Ballantine Books.

Wade, Nicholas. 1990. *Visual Allusions: Pictures of Perception*. London: Lawrence Erlbaum Associates.

Zakia, Richard D. 1979. *Perceptions and Photography*. Rochester, NY: Light Impressions.

About the Author



James M. Hoefler is assistant professor of political science and coordinator of the Policy Studies program at Dickinson College, Carlisle, Pennsylvania. Recent publications include "The Right to Die: State Courts Lead Where Legislatures Fear to Tread" (*Law & Policy* 14:4, with B. Kamoie) and *Deathright: American Culture, Medicine, Politics, and the Right to Die* (Westview Press, 1994, with B. Kamoie). He also has published articles in the areas of health care policy, state politics, and political advertising.

The Validity of Student Evaluations of Teaching

Laura I. Langbein, *American University*

Introduction and Background

Virtually all liberal arts colleges consider classroom teaching a major factor in evaluating overall faculty performance (Seldin 1989, 4). As of 1988, 80% used systematic student ratings as all or part of the means for evaluating teaching, and that percentage had increased from 68% in just five years (Seldin 1989, 4). There is also considerable agreement that systematic student ratings are reliable. Aubrecht (1981, 1), for example, reports that previous studies of student ratings, using various internal consistency measures of reliability, "show high reliabilities—in the .80s and .90s for classes of 20 or more." Similarly, Cranton and Smith (1990, 207) also report that studies of student questionnaires "generally confirm that the questionnaire is a reliable technique."

There is considerably less agreement about the validity of systematic student ratings of college teachers. Several aspects of validity have been examined, including predictive validity (Abrami, d'Apollonia, and Cohen 1990) and face validity (Aubrecht 1981, 3; Abrami, d'Apollonia, and Cohen 1990). A third aspect of validity is construct validity. Construct validity means that student ratings, if they are to be a valid measure of the quality of teaching, should be significantly associated with variables that are theoretically expected to be predictors of quality, and the ratings should not be associated with variables that are theoretically or normatively expected to be irrelevant to teaching quality. If they are associated with normatively irrelevant variables, the ratings can be said to be "biased." For example, smaller classes are expected to, and have been shown to, produce better instruction (Glass, McGaw, and Smith 1981), so if student ratings are to have construct validity, we should observe better evaluations from students in smaller classes than in larger classes when other variables are held constant. On the other hand, there is no normative reason to expect that the sex of an instructor should be related to the quality of instruction, once variables like experience, whether the course is required, and other factors are held constant. If gender and student evaluations are associated, even when other factors are held constant, the evaluations may be biased.

Previous research on construct validity has yielded inconsistent findings. The findings appear to be highly dependent on context and methodology (Abrami, d'Apollonia, and Cohen 1990; Cashin 1988), yet

https://doi.org/10.2307/420225 Published online by Cambridge University Press

evidence concerning the construct validity of student ratings of teaching is particularly important because so many colleges and universities use them to evaluate the teaching ability of college professors. This study examines the construct validity of student evaluations of teaching at the School of Public Affairs at American University. While the results may not be generalizable, the methodology is, and the larger issue of bias in student evaluations—especially gender and age bias—is clearly of general interest.

TABLE 1
Items Measuring Students' Evaluation of Teaching

1. The course was well prepared and well organized.*
2. The course materials (textbook, assigned readings, manuals, etc.) contributed significantly to my understanding of this course.*
3. The course assignments (papers, projects, homework, discussion sections, exams, etc.) contributed significantly to my understanding of this course.*
4. The course provided an appropriate amount of interaction in the classroom.*
5. Overall, the course was demanding and required high standards of performance.*
6. Overall, I am satisfied with the amount I learned in this course.*
7. Overall, this course is.**
8. The instructor's presentations were clear.*
9. The instructor was stimulating.*
10. The instructor seemed knowledgeable about the subject matter.*
11. The instructor evaluated student work carefully, impartially, objectively, and in a timely manner.*
12. Overall, the instructor is:**
13. I gained a good understanding of concepts and principles in this field.*
14. I deepened my interest in the subject matter of this course.*
15. The instructor presents recent developments in the field to the class.*
16. The instructor encourages students to think independently.*
17. Written assignments seem designed to promote the goals of this course.*
18. The instructor seems well prepared for each class.*
19. The instructor treats students with respect.*

*Response categories: Strongly agree, 1; Agree, 2; Neither agree nor disagree, 3; Disagree, 4; Strongly disagree, 5.

**Response categories: Superior, 1; Very good, 2; Good, 3; Satisfactory, 4; Fair, 5; Poor, 6.

Data and Hypotheses

At the end of each semester, American University requires faculty to give every student an opportunity to respond to a formal questionnaire that elicits closed-ended responses to 19 questions concerning the student's opinion about the course and the instructor. The questionnaire also asks other questions of fact concerning the course and the student. A School-wide personnel committee relies almost exclusively on the results of the student opinions collected by this questionnaire to evaluate the teaching performance of all faculty in the School.

The questionnaire is distributed in every class three weeks before the last class meeting. Students complete the survey during the class period. The instructor leaves the classroom while the students fill out the survey. One student places the completed surveys into an envelope, and turns the envelope in at a secure collection center. All surveys are due before the last class meeting of the semester.

This study examines data for all full-time faculty teaching in the School of Public Affairs for three semesters from spring 1991 through spring 1992. Additional information on course and instructor characteristics not included in the survey was added to the data for the purposes of this study. Because the basic purpose of this study is to examine the construct validity of student ratings, with a particular focus on finding why students rate instructors the way they do, the unit of analysis is the student, not the course. For student responses to be valid, variables that are known to reflect "better teaching" should result in better evaluations, and variables that are believed to be normatively irrelevant to teaching quality should have no impact on the evaluations. The number of usable student responses is about 2,600, though the number varies depending on missing data for the particular variables in a given analysis.

Table 1 lists the 19 questions that elicit the student's opinion about the instructor and the course. The responses to these questions constitute the dependent variable of this study. Student opinions are expected to be a function of characteristics of the student, the instructor, and the course. This study examines only the characteristics of the student that are measured by the survey instrument. As the respondent's identity is not known, it is impossible to supplement the questionnaire with other information about the student. However, the identity of the instructor and the course are known, and other data not included in the survey instrument supplemented survey information about the course and the instructor.

The survey does, however, inquire about many student characteristics thought to influence their opinions about the instructor and the course. One is their standing (Goldberg and Callahan 1991). Respondents ranged from freshmen to Ph.D. students. One expectation is that higher level students are more motivated to pursue their studies, and therefore are likely to value their instructors and their courses more than lower level students, once other variables are held constant (Aubrecht 1981). Another expectation is that upper level students are more discriminating in their evaluations and therefore will rate their instructors more critically than lower level students. Thus the sign expected for this association is not clear. The expected grade has also been found to influence opinions about the course (Aubrecht 1981; Goldberg and Callahan 1991; Scherr and Scherr 1990; Brady 1988). Students who expect high grades are hypothesized to reciprocate by giving the faculty higher evaluations. Normatively, this is thought to have a biasing impact on evaluations: faculty, believing that the expectation of a good grade will

https://doi.org/10.2307/420225 Published online by Cambridge University Press

result in better student evaluations of their teaching, respond by actually giving higher grades than they would otherwise. The result is not only an upward bias in student grades, but also an upward bias in student evaluations of the teaching of faculty who are more lenient in their grading.

Another student characteristic that might influence their evaluation of an instructor or a course is the student's actual overall grade point average (GPA). Just as higher level students are thought to be more discriminating, so also are students with a higher GPA thought to be more discriminating in their evaluations of teachers and courses. However, the expected sign of the association between GPA and the student's rating of the instructor and the course is not clear. If the rating reflects the quality of teaching, then better students can be expected to give better ratings to better courses and instructors. By contrast, if the rating measures how entertaining the instructor is, in the absence of substance, then one would expect that better students will award entertainers with poorer ratings.

Two other student characteristics thought to influence their evaluations include the number of hours (including class and laboratory time) that they spent on the course, and the number of times they consulted with the instructor outside of class. Students who have spent more time on the course, or meeting with the instructor outside the class, are likely to rate the instructor and the course more highly at the end of the semester than those who have spent little time on the course, even when other variables are held constant. Normatively, it is particularly important that the amount of time students spend on the class result in higher evaluations of an instructor. If the ratings are a valid measure of teaching quality, and if high quality teachers are those who have motivated the student to spend time on the course material, then a significant association between these variables, even when other factors are held constant, would uphold the construct validity of the ratings (Prosser and Trigwell 1990; Brady 1988).

Several course characteristics are often thought to influence student evaluations. One is whether the course is required (Cashin 1988; Scherr and Scherr 1990). Courses can be required by the university, or by the major, the minor, or a program. Because students choose their major, their minor, and their program (e.g., the MPA program), courses required as a result of that choice are not likely to be regarded as a burden in the same way that courses required of all undergraduates in the university are regarded. Hence, courses required by the university were coded "1" and other courses were coded "0."[1] Another characteristic of the class is its size. Previous research has shown that smaller classes produce higher levels of student achievement in the course, especially when other variables are controlled (Glass, McGaw, and Smith 1981). Therefore, if, in the results below, smaller classes result in better evaluations of the instructor and the course, this would uphold the construct validity of student evaluations as a measure of the quality of instruction.

A third characteristic of the course that is particularly relevant for social science programs is whether the course is quantitative or not. All undergraduate and graduate statistics and research methods courses taught by the School of Public Affairs were coded "1," and other courses were coded "0." Quantitative courses are expected to be rated lower than other courses, especially when other variables are held constant. The final course characteristic examined by this study is whether the course is "nontraditional." The School of Public Affairs (SPA) offers weekend courses to students in some of its master's degree programs and intensive 1-week special courses to undergraduates who choose them. Because these courses meet for entire days rather than for a few hours at a time over a semester, there is considerable interaction among the students and between the students and the instructor. The master's level courses are also tailored for the specific class. This special format makes it reasonable to expect that student ratings of the course and the instructor will be higher for these courses than for the "regular" courses, even when other variables are held constant.

Finally, characteristics of the instructor are also believed to influence student ratings of both the instructor and the course (Kierstad, D'Agostino, and Dill 1988; Smith and Kinney 1992; Goldberg and Callahan 1991; Ghorpade and Lackritz 1991; Basow and Silberg 1987). This study examined two faculty characteristics. One is experience. In this study, experience is measured as the number of years of full-time teaching, either at AU or elsewhere. Experience is expected to have a nonlinear relation with student ratings. Inexperienced faculty probably begin with low ratings; the ratings improve with more experience, but then level off or decline as a result of the growing age disparity between the instructor and the students, as a result of "burnout," or simply because the better one's ratings are, the harder it is to improve them further. The other faculty variable is gender. Gender is expected to have a variety of effects on ratings, even when other variables are held constant. First, one possibility is that gender has a direct effect on ratings. Students may discriminate against female faculty, just as the larger society discriminates against women in the workplace, and may consistently give them lower ratings than comparable male colleagues.

Conversely, it is also possible that students give female faculty higher ratings; if women really are "warmer" and more nurturing toward students, as their stereotype would suggest, and if being "warm" and "nurturing" elicits higher evaluations, then female faculty could get higher ratings than male faculty. It is also possible that the female faculty of the School are genuinely better (or poorer) instructors than their male colleagues. Any of these expectations implies that there will be significant association between gender and ratings, even when other variables are held constant. However, there is no rea-

https://doi.org/10.2307/420225 Published online by Cambridge University Press

son to expect a particular direction; and the interpretation of a significant association is not clear. For example, if women are rated significantly worse than men, even when other variables, including experience and expected grade, are held constant, one cannot conclude that the ratings are biased. Female instructors could be rated more poorly than their male colleagues because students discriminate against women; but they could also be rated more poorly simply because, at SPA, they really are worse as instructors than their male colleagues. Thus, a statistically significant direct effect would not necessarily be an indicator that student ratings are biased.

However, the theory of discrimination has some other testable implications that the theories implying direct effects do not share. Specifically, both theory and previous empirical evidence support the argument that there are conflicting expectations about professional women in the workplace (Kierstad, D'Agostino, and Dill 1988; Martin 1984). On one hand, women are expected by their students to have "feminine" characteristics; this means they are to be warm, friendly, and supportive. On the other, they are expected by their peers to do their job; this requires "masculine" characteristics, and women are expected to be objectively critical, to be assertive, and to exercise authority when necessary. In the context of the classroom, this implies that when a student expects a low grade from a female faculty member, the student will give the female faculty member a lower evaluation than a comparable male faculty member would get. That is, the theory of discrimination implies that gender and expected grade will have an interactive effect on the student's rating of the instructor. When being authoritative (the student expects a low grade) is inconsistent with a traditional role expectation (women are supposed always to be supportive), the result is a low rating from students.

Another consequence of incommensurate role expectations that society has of professional women in the workforce is that students will give female faculty with whom they spend less time outside of the class a lower rating than a comparable male with whom they spend the same amount of time outside the class. In other words, the impact of time spent with the instructor outside the class is expected to raise student ratings more for female faculty than for male faculty. Spending lots of time with a student outside of class is "nurturing" and role consistent for female faculty, who will be rewarded with higher ratings than comparable male faculty; spending relatively less time with a student outside of class is perceived as being less supportive, and therefore as role inconsistent for female faculty, who will be punished with lower ratings than comparable male faculty.

Both instances of statistical interaction are expected if students rate men and women faculty differently. It follows that if the parameter estimates associated with one or both of these interaction terms are statistically significant and have the expected sign, this would provide evidence that student ratings have a source of bias.

Findings

Table 1 sets forth the 19 items on the student evaluation of teaching (SET) that ask for the student's opinion about the course and the instructor. One might hypothesize that questions pertaining to the course tap a different dimension than items referring to the instructor, or that there are other aspects of multidimensionality in how students evaluate teaching, such as the instructor's presentation skills, ability to facilitate learning, and adeptness at classroom management (Marsh 1991). However, for students in SPA responding to the survey items listed in Table 1, Table 2 reveals a clear-cut single dimension. The principal components analysis extracted only one factor with an eigenvalue greater than unity; the eigenvalue of that first vector is 10.2, while that of the second axis drops sharply to .95. The first principal axis explains 54% of the total variance in the 19 items. Moreover, all 19 items correlate substantially and about equally with that single vector. The facts that the principal component has such a large eigenvalue, that the remaining factors have considerably smaller eigenvalues all below the arbitrary but common criterion of unity, and that all the items load nearly equally and predominantly on the first factor, support the conclusion that SPA's student evaluations of teaching are unidimensional.

TABLE 2
Factor Pattern Matrix, 19 SET Items

Principal Factor	Variable
.77	Course well-prepared (Q1)
.61	Course materials (Q2)
.73	Course assignments (Q3)
.69	Interaction (Q4)
.59	Course demanding (Q5)
.84	Satisfied with amount learned (Q6)
.89	Overall course (Q7)
.80	Instructor presentations clear (Q8)
.81	Instructor stimulating (Q9)
.67	Instructor knowledgeable (Q10)
.68	Instructor careful, etc. (Q11)
.76	Overall instructor (Q12)
.83	Understanding of concepts (Q13)
.75	Deepened interest (Q14)
.65	Recent developments (Q15)
.70	Encourage to think independently (Q16)
.68	Written assignments (Q17)
.73	Instructor well-prepared (Q18)
.68	Treat with respect (Q19)
Eigenvalue 10.21	

Unidimensionality makes it possible to create a Likert scale by summing the scores for all 19 questions, so that each student gives a total rating score to each course. The scores for each response category listed in Table 1 were reversed, so that a score of 19 means the student rated the course at the bottom on all 19 items; a score of 97 means the student rated the course tops on all 19 questions.[2]

This scale is used as the dependent variable in the analyses reported below. Most scores are quite high, indicating a positive rating. Only 25% of the responses are below the numerical midpoint of 39.5, and 90% of the observations are above 50. The highest score (97) is also the modal score. Based

https://doi.org/10.2307/420225 Published online by Cambridge University Press

on Cronbach's alpha, the scale's reliability is 0.99. By definition, the reliability of a multi-item scale exceeds that of any single-item scale. Nonetheless, all of the results reported below are also examined using as a dependent variable not only the 19-item scale but also the two single questions that SPA relies on for its personnel decisions—item Q7, which asks for an overall assessment of the course, and item Q12, which asks for an overall assessment of the instructor. In general, as Table 3 reveals, the findings for each dependent variable are the same.

Table 3 shows that nearly all the student characteristics hypothesized to influence their evaluation of teaching in fact have an impact. For instance, Table 3 shows that students with a higher class standing give significantly lower scale scores and lower overall course scores, indicating a lower total evaluation and a lower overall course evaluation. However; there is no significant association between a student's standing and the single-item response rating the instructor overall. These results suggest that when other factors are held constant, upper level students are more discriminating or critical in their evaluation of courses.

As hypothesized, the expected grade in the course has a significant effect in the predicted direction, no matter whether the rating is measured as the totality of the student's responses or as a single item measuring the student's opinion about the class and the instructor overall. For each additional unit increase in the expected grade (e.g., B to A, or C to B), the total scale score increases by 3.36 points, indicating that a better evaluation is associated with a higher expected grade, even when other variables are held constant.[3]

Compared with the expected grade, the student's actual overall GPA has an opposite but not quite offsetting impact on evaluation. For each additional unit increase in actual GPA (e.g., from 3.0 to 4.0), the total scale score decreases by 1.00, indicating a poorer evaluation. This association between GPA and rating remains significant and

TABLE 3
Regression of Total Scale Score, Overall Instructor, and Overall Course on Independent Variables

Independent Variable	Total Scale Score	Overall Instructor	Overall Course
Intercept			
Param. est.	78.12****	5.16****	4.52****
T-stat.	7.34	3.82	5.33
Stdized est.	—	—	—
Student level			
Param. est.	−.65***	−.02	−.07****
T-stat.	−3.29	−.91	−3.86
Stdized est.	−.07	−.02	−.09
Expected grade			
Param. est.	3.36****	.23****	.29****
T-stat.	7.63	5.55	7.32
Stdized est.	.18	.13	.17
Grade point avg.			
Param. est.	−1.00**	−.09**	−.07*
T-stat.	−3.27	−3.23	−2.49
Stdized est.	−.07	−.07	−.05
Hours spent			
Param. est.	2.23****	.08**	.19****
T-stat.	8.49	3.10	8.23
Stdized est.	.16	.06	.16
Times consulted w/inst.			
Param. est.	1.58***	.23****	.09*
T-stat.	3.75	5.56	2.43
Stdized est.	.08	.13	.05
Required course			
Param. est.	−.23	−.05	−.05
T-stat.	−.43	−.93	−.99
Stdized est.	−.01	−.02	−.02
Class size			
Param. est.	−.06**	.00	−.005*
T-stat.	−2.59	.00	−2.11
Stdized est.	−.05	.00	−.04
Quantitative course			
Param. est.	−1.00	−.11	−.18*
T-stat.	−1.10	−1.37	−2.25
Stdized est.	−.02	−.03	−.04
Non-traditional course			
Param. est.	4.58*	.31	.60***
T-stat.	2.56	1.88	3.72
Stdized est.	.05	.04	.07
Yrs. of experience			
Param. est.	.78****	.07****	.07****
T-stat.	7.50	7.03	6.98
Stdized est.	.58	.57	.54
Experience squared			
Param. est.	−.03****	−.002****	−.002****
T-stat.	−8.35	−7.99	−7.53
Stdized est.	−.63	−.62	−.57
Sex (M = 0; F = 1)			
Param. est.	−13.89****	−1.52****	−1.15***
T-stat.	−3.93	−4.62	−3.60
Stdized est.	−.52	−.62	−.48
Sex ∗ expected grade			
Param. est.	2.29**	.23**	.15*
T-stat.	3.11	3.42	2.33
Stdized est.	.38	.42	.28
Sex ∗ times consulted			
Param. est.	.46	.13	.08
T-stat.	.61	1.87	1.15
Stdized est.	.03	.11	.07
Adjusted R^2	.12	.10	.11
F-stat	28.22	21.13	25.62
Number of obs.	2760	2760	2760

*Prob. ≤ .05.
**Prob. ≤ .01.
***Prob. ≤ .001.
****Prob. ≤ .0001.

https://doi.org/10.2307/420225 Published online by Cambridge University Press

positive no matter whether the rating is measured as a scale or as a response to a single item.[4] Had the association been positive, it would have indicated that better students evaluate instructors and courses with more approbation; such a result would be consistent with what would be expected if the responses in the SET were truly reflective of the quality of teaching. However, the association is not positive. Rather, the negative association implies that better students are more disapproving about their courses and instructors and possibly more discriminating in their judgments so that they do not reward performance without substance. Conjointly, students who expect a high grade for the class reward the instructor with a more positive evaluation. Thus, students with a proven track record (a high GPA) give their instructors a lower evaluation, while those who merely expect to do well in a given class respond by giving a higher evaluation. Together, these findings do little to support the argument that a high score on the SET is truly a measure of top-quality teaching.

However, the opposite conclusion emerges from the findings about time spent on the course and with the instructor. The results in Table 3 support the expectation that students who spend more time on the course, or with the instructor after class, rate the course better at the end of the term. For example, each additional two hours spent on the course results in a scale score increase of 2.23 points—a higher evaluation. Similarly, consulting more with the instructor outside of class is associated with a scale score increase of 1.58 points—again, a higher evaluation. Both associations retain their sign and their statistical significance even when the dependent variable is the single-item overall evaluation of the course or the instructor. The fact that students who spend more time on a class also regard it more highly is particularly important. Many believe that the best teachers are those who inspire students to spend more time on the subject they teach. The results in Table 3 suggest that this wish has some empirical grounding, and further imply that at least some aspect of the SET does tap a portion of a measure of the quality of teaching.

Table 3 also reveals that characteristics of the course, overall, have a mixed impact on course ratings, once other variables are held constant. For example, courses that are required by the university are not rated differently from other courses, no matter whether the rating is measured by the total scale score or by the response to the two single items asking about the overall evaluation of the course and the instructor.

Previous research has suggested that students learn more in smaller classes. If the SET score is indicative of instructional quality, then lower scores (indicative of poorer evaluations) should be significantly associated with larger classes. Table 3 supports this expectation for the total scale score and for the overall course rating, but not for the instructor rating. The parameter estimate for the overall scale score means that every additional 10 students reduces the evaluation by 0.6. While the magnitude of the impact is not large, the direction is consistent with what would be expected if the SET were to be construed as a measure of quality.

Social science students usually dread quantitative courses. Table 3 reveals that, of the three dependent variables, only the single-item course rating is significantly related to whether the course is quantitative, once other variables are held constant. Quantitative courses score lower on the single-item response scale (that is, they are evaluated more poorly). But, when the overall scale score is examined and other variables are controlled, they are evaluated no differently than nonquantitative courses.

Because of their special format, "nontraditional" courses were expected to be evaluated more positively than other courses, and Table 3 upholds this expectation. For example, nontraditional courses score 4.58 points higher on the total scale score than otherwise comparable regular courses, and the positive association is also significant for the single-item course and instructor ratings (especially with a 1-tailed test, which is not inappropriate in this case, as the positive association was expected by theory).

Finally, Table 3 reveals that characteristics of the instructor affect not only the total scale scores and the single-item rating of the instructor, but also the single-item rating of the course. Specifically, the number of years of full-time teaching experience at American University or another university has a significant nonlinear impact on those dependent variables. For example, based on the results for the total scale, evaluations become more positive during each of the first 13 years of teaching experience; then the scores turn downward. A similar pattern emerges for the single item responses.[5] There are increasingly positive evaluations as years of teaching experience increase, but the pattern switches in the mid-teen years of experience, when more experience (or, more probably, age or burnout) turns student evaluations in a negative direction.

Many (but not all) previous studies have shown that female instructors are rated more poorly by students than comparable male instructors. This study lends partial support to these findings. No matter whether the rating is measured by the total scale score or by the response to a single item rating the course or the instructor, this study shows that the effect of gender is not a simple additive one. Rather, the impact of gender interacts with expected grade but not with the number of times a student consults with the instructor outside the class.

Consider first the interaction of gender with expected grade. Specifically, based on the total scale scores, the results in Table 3 imply that if all other variables were held constant at zero, the regression equation for men would be

$$Y' = 78 + 3.36 \text{ (expected grade)}$$

while that for women would be

$$Y' = 64.11 + 5.65 \text{ (expected grade).}$$

This means that if a male faculty member were to raise an expected grade by 1 point, his total score

https://doi.org/10.2307/420225 Published online by Cambridge University Press

would rise by 3.36 points. If a female faculty member were to take the same "supportive," "nurturing" step, her evaluation would go up by about 2 points more than that for men—by 5.65 points. Conversely, for women, as expected grades decrease, their evaluation falls by more than that for comparable male faculty. Thus, female faculty are rewarded, relative to men, for "supportive," "nurturing" behavior, but they are punished, relative to men, for "objective," "authoritarian" behavior that is role inconsistent. In addition, the modal expected grade in SPA is just between an A and a B, scored at 4.5 on the survey expected grade scale. The regression equation for women implies that female faculty from whom this grade is expected score 89.54 on the total scale, whereas otherwise comparable male faculty score 93.12, or about 3.5 points more, which is a better evaluation. Now, suppose that the expected grade were a "C," scored 3.0 on the expected grade scale in the survey. In that case, a male faculty member who scored zero on all other variables besides expected grade could expect to be scored 88.08 on the total rating scale, while a comparable female faculty member could expect a score of 81.06 on the same scale—a seven point lower evaluation. On the other hand, when A is the expected grade, female faculty are still rated more poorly than comparable male colleagues, but the difference is less—only some 2.5 points.

Overall, these results support the theoretical expectation that students treat female faculty members differently from otherwise comparable male faculty members. When a student expects a low grade from a female faculty member, this is regarded as not "nurturing" and the female instructor receives a lower evaluation than her comparable male colleague would. The lower the expected grade, the more poorly the female instructor is evaluated relative to a comparable male colleague.

Table 3 also shows little support for the expectation that students who consult many times with a female instructor outside class will reward that "role consistent" instructor with a higher evaluation than an otherwise comparable male instructor would receive. Specifically, the interaction between gender and the number of times the student consulted with the instructor outside class has no statistically significant impact on any of the three dependent variables.

Table 3 displays another important aspect of student evaluations: most of the variance is unexplained by the 14 student, course, and instructor characteristics included in the statistical model. No matter whether measured as a total scale or as a response to a single item rating the course or the instructor, some 88% to 90% of the variance in student ratings is unexplained. Nor does the goodness of fit improve when a variety of nonlinear functional forms was examined. For example, when the log of the total scale was used as a dependent variable, the R^2 was 0.13. When a multinomial logistic regression of the two single-item responses was used to estimate the parameters of the model, the pseudo-R^2 was less than .10 for both items (Aldrich and Nelson 1984, 57).

While all the goodness of fit significance tests (that is, the F-test for the OLS regression and the X^2 test for the logistic regressions) indicate that including the 14 independent variables improves the model compared with using the intercept alone to predict the dependent variable, the magnitude of the goodness of fit statistics is not high. This means that many variables may be omitted from the model, or that there is considerable randomness in student ratings of teaching. The facts that the total scale has a reliability of 0.99 and that it appears from a factor analysis to be unidimensional imply that random measurement error is not likely to be a major source of the unexplained variance. Previous research suggests several student characteristics that should be included. One is the gender of the student or the gender mix in the class, although the research findings here are extremely mixed (Abrami, d'Apollonia and Cohen 1990; Kierstad, D'Agostino, and Dill 1988; Cashin 1988). Another is student personality, or the "match" between the personality of the instructor and the student (Abrami, d'Apollonia, and Cohen 1990; Cashin 1988). There is no reason to believe that these omitted variables are related both to the dependent variables in this study and to any of the included independent variables, so their omission is not likely to mean that the parameter estimates reported in Table 3 are biased.

Summary and Implications

Overall, based on both the unstandardized as well as the standardized parameter estimates, the results in Table 3 show that, of the variables examined, course characteristics have the smallest impact on student ratings, student characteristics have a mid-range impact, and the faculty characteristics of gender and experience have clearly the largest impact. But the effect of experience is not linear, and that of gender interacts with other variables, so one cannot conclude that more experience always contributes to better evaluations or that the extent to which female faculty members are evaluated more harshly than comparable male colleagues remains constant. The substantive impact of even the most important variables is not large, nor is the total explained variance.

It is, in fact, unclear exactly what the student ratings really measure. There is evidence to support the argument that the ratings are a popularity contest as well as the argument that the ratings are a measure of quality instruction. Supporting the former argument, the fact that a higher expected grade consistently and significantly raises student opinions of instructors implies that students favor instructors from whom they expect good grades, no matter how hard the student works. Moreover, the fact that better students—those with a high GPA—rate instructors more negatively implies that the students who are best able to separate good entertainment and easy grading from real learning do just that. When the

https://doi.org/10.2307/420225 Published online by Cambridge University Press

truly best students give instructors the poorest ratings, while the students who merely expect good grades give the same instructors the highest evaluations, the implication is that the ratings measure how easy the course is, not how much the student is learning.

Other evidence, however, suggests that the ratings reflect the quality of instruction. For example, students who spend more total time on the course and more time consulting with the instructor outside the class also give more positive ratings. This implies that the ratings really reflect at least some of what quality instruction is supposed to do. Instructors who can motivate students to work hard on the material they teach surely deserve a favorable evaluation. Also, previous evidence suggests that students learn more in small classes. The evidence in this study generally shows that smaller classes get better evaluations. If this occurs because more learning comes about in smaller classes, then the associations would be consistent with the argument that SETs measure the quality of teaching.

Based on the results of this study, it is best to conclude that mostly we do not know what SETs measure. Partly they reflect quality; partly they reflect popularity and "gut" courses; and in even greater part they reflect the age and gender of the instructor—but not in ways that are linear and additive, respectively.

Until we learn more about what students' responses to systematic surveys of teaching really measure, it is probably unwise to rely exclusively or even predominantly on them as a means of comparing the teaching performance of one faculty member to that of another. We currently rely heavily on student surveys because they appear to be a relatively costless means of evaluating the teaching effectiveness of faculty, compared with other means—such as actually measuring what students learned in the course, directly observing lectures and discussions, or surveying alumni, faculty peers, and university administrators. But the opportunity costs of SETs could be quite high. To the extent that SETs reward poor quality, gut courses, and entertainment without substance, their cost, though hidden, could in fact be higher than the cost of any of the alternatives. It is also possible that current SETs are a cost-effective means of gauging teaching effectiveness. Based on the studies available, there is no reason to believe that the findings from one college setting will apply in another.

At this point, SETs are a tool—but we don't know what it does, when it works, and when it doesn't. It is probably a good time both to find out what SETs really measure and to supplement the SETs with other, related tools that share similar strengths but have different weaknesses. It is not reasonable to expect that a single methodology will measure teaching quality with reliability and validity. SETs appear to be reliable but to have questionable validity. Other measures, such as direct observation, may have dubious reliability but considerable validity. Together, however, the use of multiple measures makes it possible to attain a reasonable degree of both reliability and validity.

Notes

[1] Recoding this variable so that courses required by the major or minor were also regarded as being "required" had no effect on the results reported below.

[2] Once the scores in Table 1 are reversed, for 17 items, a "5" is the highest rating; for 2 items, a "6" is the highest rating. Hence (5 × 17) + (2 × 6) = 97 is the highest possible score and the highest possible rating.

[3] While this study cannot conclusively rule out the possibility of reciprocal causation between student evaluations and expected grade, results using an appropriate estimation procedure were inferior to those presented in Table 3.

[4] Tests showed no evidence that SPA instructors "teach toward the middle"; at least, students in the middle of the GPA distribution do not respond with higher evaluations.

[5] Tests showed that ratings rise and fall with experience, rather than merely level off; moreover, the nonlinear relation is not due to a few extreme scores on the experience variable.

References

Abrami, Philip C., Sylvia d'Apollonia, and Peter A. Cohen. 1990. "Validity of Student Ratings of Instruction: What We Know and What We Do Not." *Journal of Educational Psychology* 82(2):219–231.

Aldrich, John H., and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-045. Beverly Hills, CA: Sage.

Aubrecht, Judith A. 1981. "Reliability, Validity, and Generalizability of Student Ratings of Instruction." *Idea Paper 6*. Center for Faculty Evaluation and Development, Kansas State University (November).

Basow, Susan A., and Nancy T. Silberg. 1987. "Student Evaluations of College Professors: Are Female and Male Professors Rated Differently?" *Journal of Educational Psychology* 79(3):308–314.

Brady, Peter J. 1988. "The Effects of Course Demands and Grades on Anonymous versus Nonanonymous Evaluations of Professors." Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans (April).

Cashin, William E. 1988. "Student Ratings of Teaching: A Summary of the Research." *Idea Paper 20*. Center for Faculty Evaluation and Development, Kansas State University (September).

Cranton, Patricia, and Ronald A. Smith. 1990. "Reconsidering the Unit of Analysis: A Model of Student Ratings of Instruction." *Journal of Educational Psychology* 82(2):207–212.

DeCanio, Stephen J. 1986. "Student Evaluation of Teaching—A Multinomial Logit Approach." *Economic Education* 17(3): 165–176.

Ghorpade, Jai, and James R. Lackritz. 1991. "Student Evaluations: Equal Opportunity Concerns." *Thought and Action* 7(1):61–72.

Glass, G.V., B. McGaw, and M.L. Smith. 1981. *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage.

Goldberg, Gerald, and John Callahan. 1991. "Objectivity of Student Evaluations of Instructors." *Journal of Education for Business* 66(6):377–378.

Kierstad, Diane, Patti D'Agostino, and Heidi Dill. 1988. "Sex Role Stereotyping of College Professors: Bias in Student Ratings of Instructors." *Journal of Educational Psychology* 80(3):342–344.

Marsh, Herbert W. 1991. "Multidimensional Students' Evaluations of Teaching Effectiveness: A Test of Alternative Higher-Order Structures." *Journal of Educational Psychology* 83(2):285–296.

Martin, Elaine. 1984. "Power and Authority in the Classroom: Sexist Stereotyping in Teaching Evaluations." *Journal of Women in Culture and Society* 9(3):482–492.

Prosser, Michael, and Keith Trigwell. 1990. "Student Evaluations of Teaching and Courses: Student Study Strategies as a

https://doi.org/10.2307/420225 Published online by Cambridge University Press

Criterion of Validity." *Higher Education* 20(2):135–142.

Seldin, Peter. 1989. "How Colleges Evaluate Professors." *American Association for Higher Education Bulletin* 41(7)(March):3–7.

Scherr, Frederick C., and Susan S. Scherr. 1990. "Bias in Student Evaluations of Teacher Effectiveness." *Journal of Education for Business* 65(8):356–358.

Smith, Sharon P., and Daniel P. Kinney. 1992. "Age and Teaching Performance." *Journal of Higher Education* 63(3):282–302.

About the Author

Laura I. Langbein is a professor in the School of Public Affairs at American University, where she teaches quantitative methods and policy analysis. She has published articles on methodology, the implementation of environmental regulations and, most recently, the impact of campaign contributions and lobbying on legislative voting.

Teaching Research Methods Using Appropriate Technology*

Allan McBride, *Grambling State University*

Undergraduate and graduate-level courses in social science research methods are widely avoided and maligned by students while faculty members who are required, or who choose, to teach these courses often suffer from poor student evaluations. The reasons for this situation are related to the nature of the material, which leaves little opportunity for students to apply the knowledge they have gained in other courses in their major; at least students believe this to be the case. Additionally, students are required to master the language of scientific methods, with specific and technical definitions; to understand scientific and experimental notation; and to comprehend the difficult area of probability theory and its relation to sampling, all matters for which students see little or no purpose. Yet faculty recognize that students who are to be well-informed citizens, or to attend graduate school, or to seek professional employment upon graduation need to master some of these skills to contribute successfully to their communities and to their professional lives.

Further contributing to this disturbing scenario is the evidence that many undergraduate, and possibly some graduate, students lack cognitive sophistication. Hudak and Anderson (1990) report that as many as 50% of undergraduate students are not capable of abstract thinking. Whether this is evidence of a flawed theory of human development, a failure of secondary and elementary education, or the result of more widely available postsecondary education, is unclear. However, it does suggest that university faculty must be more sensitive to students' limitations and design their courses to take advantage of the capabilities that they have when they arrive on campus.

I have taught research methods for eight years at an open admissions HBCU (Historically Black College/University) both at the graduate and undergraduate levels and have had to deal with all the problems mentioned above, writ large in comparison to faculty who have the luxury of teaching in more selective universities. Even so, I think it is possible to engage students in the research process, even in a single-semester course, so that they can experience the pleasure and excitement of conducting research while learning its basic principles.

It is my purpose in writing to share my experiences with other faculty concerning some practical, hands-on methods for teaching a course in research methods—methods that are suitable for students who are at what Piaget referred to as the concrete operations stage of development. The approaches that I discuss in this paper are suitable for both undergraduate and graduate classes, and can be used successfully even where student research sophistication is quite low.

Appropriate Technology for Research Methods

The term "appropriate technology" was coined in the 1970s to describe energy generation methodologies that were readily accessible to a broad sector of the population, were relatively inexpensive, and were easy to employ. I use the term in the same sense to suggest that the methods of research can be accessible, inexpensive, and easy to use, particularly with the development of personal computers (though not solely for that reason). Below I identify some approaches to research that meet these three criteria.

*The ability to conduct the type of research described in this paper has been enhanced by National Science Foundation Instrumentation and Laboratory Enhancement grant #USE-9251254. Grants of this sort are available to undergraduate institutions to help them improve laboratory facilities for undergraduate uses.

https://doi.org/10.2307/420225 Published online by Cambridge University Press