34 What every economist should know about the evaluation of teaching: a review of the literature

Stephen B. DeLoach

The subject of teaching evaluations is uniformly controversial within academe. Despite decades of research verifying their reliability, their validity continues to be questioned by researchers and faculty. One reason is that such methods only consider a relatively small number of factors that correlate with "good instruction." Even though these limitations are well-established in the literature and widely acknowledged by faculty, the implementation of alternative or complimentary forms of assessment is far from universal.

What students are capable of assessing does not always coincide with what is required for learning. Evaluation of teaching and evaluation of learning are two different things (Becker and Watts, 1999). Evidence strongly suggests that students can reliably assess factors such as teaching skill, rapport with students, organization, difficulty, and feedback (Paulsen, 2002). They cannot assess subject mastery, curriculum development and course design (Arreola, 2000; Chism, 1999).

The literature also indicates that student evaluations of teaching are biased. Four factors are routinely cited as primary sources for these biases. These include prior interest in subject, expected grades, perceived workload, and reasons for taking course (Marsh, 1987). Interestingly, recent research in economics by Weinberg, Hashimoto and Fleisher (2009) shows that these biases may also have disparate effects on different courses. The implication is that one-sized-fits-all evaluations are inappropriate.

The vast majority of existing methods of evaluating instruction are mandated by administrators and designed at the university level (Becker and Watts, 1999). Nevertheless, it is difficult to argue that good economics instruction is the same as good biology or language instruction. As a result, university-administered student evaluations of teaching are likely to capture only a small portion of what we might believe is relevant for economics. With all the controversy over the use of student evaluations to assess effective teaching, one might think that economics departments have developed more sophisticated instruments to use in assessing instruction. While alternative methods exist, departments have been slow to implement anything but the most superficial of these alternatives.

The purpose of this chapter is to review the current methods used to assess teaching. While evaluations can be used either for summative or formative purposes, the evidence strongly suggests that most departments are primarily concerned with summative feedback. This will be the primary focus of this chapter. Different methods for evaluating teaching are reviewed, including evidence of their prevalence across both economics and the academy in general, with special attention paid to research on the reliability and validity of alternative assessments. This is followed by a description of the use of assessments, including both summative and formative uses. The chapter concludes with some recommendations for the discipline.

REVIEW OF CURRENT METHODS FOR ASSESSING TEACHING

Fundamentally, there are two ways of formally assessing instruction: (1) student end-of-course evaluations; and (2) peer evaluation.[1] The choice of method by departments and universities depends on a number of factors, including the purpose of the evaluation (in other words, annual merit raises, promotion and tenure, etc.), the size of the institution, and the mission of the school or department (in other words, liberal arts colleges, research universities, PhD programs, etc.).

Student Evaluation

The most common method of evaluating teaching, even at liberal arts colleges, is end-of-course student evaluations. These evaluations typically consist of fixed-response questions scored on a Likert scale (Cashin, 1999; Seldin, 1999), complemented by open-ended questions where students can write their opinions of the instructor's relative strengths and weaknesses. In their 1998 survey, Becker and Watts (1999) found that 83 percent of economics departments used fixed-response qualitative measurement instruments with open-ended questions on forms. Areas most often assessed include overall effectiveness, communication skills, organization, and knowledge of material; the least likely are decorum, use of technology, rapport and use of applications and examples.

Research suggests that such evaluations have high inter-rater reliability.[2] Centra (1993) notes that, as long as more than ten students are surveyed, inter-rater reliability is generally high (0.70 and higher). Reliability over time is also good, with inter-rater reliability of at least 0.83 (Marsh and Dunkin, 1997). Reliability, however, is not the same as validity. In general, metrics must meet both criteria to be considered good measures of teaching effectiveness.

One way of assessing validity is how correlated student measures are with other forms of evaluation, referred to as "construct validity" (Marsh, 2007). Here, the evidence appears to support validity. For example Feldman (1989a) finds that students' end-of-course evaluations are correlated with alumni (0.69), instructors (0.29), colleagues (0.55), administrators (0.39), and external, trained evaluators (0.50). Despite the demonstrated construct validity, this still falls short of validating whether student evaluations are correlated with student learning.

Validation of learning experiments typically employs multiple sections of a course. Not surprisingly, a number of tight controls are needed to make these experiments well-designed.[3] Sections must use the same textbook, variation in student enrollment must be controlled (through either randomized assignment or econometric sample selection controls), and consistent pre- and post-tests must be administered. In an analysis of 41 suitably designed studies, Cohen (1987) found that student achievement was positively correlated with factors typically evaluated by students on end-of-course assessments. For example, learning was correlated with student assessments on factors such as course structure (0.55), interaction (0.52), instructor skill (0.50), overall course (0.49), overall instructor (0.45), rapport (0.32), and feedback (0.28).

A great deal of evidence shows measures of teaching effectiveness are biased by a number of factors, including prior interest in the subject, expected grades, perceived

workload, and reasons for taking course (Marsh, 1987). However, not all biases are evidence of a lack of validity (Marsh, 2007). In fact, many of the factors affecting student evaluations *should* affect both learning and teacher effectiveness. Examples of these include class size and prior interest in the subject. Not surprisingly, the most controversial bias factor is grades.

A number of studies find evidence that grading leniency increases student end-of-course evaluations, but the effect appears to be relatively small (Marsh 2007). Recent evidence in economics is more damning. In a study from principles sections at Ohio State University from 1995–2004, Weinberg et al. (2009) find that a student's current grade is a significantly positive determinant of evaluation scores. In contrast to earlier studies, Weinberg et al. (2009) find an extremely large effect of grades on evaluations. Interestingly, the grade-induced bias is roughly triple for macroeconomics than for microeconomics courses. Moreover, they find that after controlling for current grade, learning is not significantly correlated with student evaluation scores.

With the growing use of online methods for administering student evaluations, many fear that sample selection bias has further eroded the validity and reliability of student evaluations. However, the existing evidence appears to refute this. The obvious advantage in such systems has to do with the cost and ease of implementation. One unexpected advantage in online evaluations is that students appear to write longer, more detailed comments (Hardy, 2003). The main concern in administering evaluations online is the response rate. As expected, response rates are lower with web-based evaluations than with traditional methods (Hardy, 2003; Avery et al., 2006). While this increases the standard error of point estimates, there is little evidence that mean ratings are significantly affected. In a recent study in a large economics-based public policy school, Avery et al. (2006) found virtually no difference in mean responses to any of the survey questions. They did, however, find evidence that *who responds* differs online. Females and those expecting higher grades were significantly more likely to respond. Their evidence also suggests that response rates increase over time once web-based systems are implemented.

As serious as the concerns over the validity of student evaluations is the criticism that they only assess a limited set of factors related to good teaching (Becker, 2000). Even if valid, they represent a narrow range of factors generally accepted as necessary for effective teaching. According to Cashin (1999), students are in a position to evaluate delivery of instruction, clarity of presentation, availability to students and administrative requirements. They cannot evaluate subject mastery, curriculum development and course design, factors appropriate for peer evaluation (Arreola, 2000; Chism, 1999).

Peer Review

Despite widespread acknowledgment that it is needed, peer review is significantly less prevalent than student-based evaluations of teaching. At a minimum, peer assessment is needed to evaluate course content. Since faculty members' work is valued more highly when it is subjected to rigorous review, peer review should result in the increased importance of teaching in the overall evaluation of faculty (Chism, 1999). The most commonly employed method is direct classroom observation, although the use of teaching portfolios is increasing.

According to Seldin (1999), undergraduate institutions – liberal arts colleges and small universities – are the most likely to use peer review of teaching. Collecting data from liberal arts colleges over a twenty-year period from 1978–98, Seldin (1999) reports that, as early as 1978, nearly half of all liberal arts colleges claimed to require peer review in the form of committee review, even more by the Dean and department chair. However, in 1978 less than 15 percent of liberal arts colleges mandated either classroom visits or review of teaching materials by peers as part of the review process. By 1998 this figure had grown to over 40 percent. Interestingly, nearly 10 percent of colleges also report that analysis of grade distributions is part of the review process. Overall, these findings suggest that the evaluation of teaching is becoming more multifaceted and sophisticated.

Seldin's findings are consistent with what appears to be going on in economics departments as well. White (1995) found that 26 percent of departments required classroom visits and 30 percent reviewed teaching materials; 59 percent used multiple methods, including formal and informal follow-ups. In a later, more comprehensive survey, Becker and Watts (1999) found that 52 percent of Bachelor, 51 percent of Master's, 42 percent of Doctoral, and 37 percent of Research institutions required some form of peer review. While classroom observation was most frequently cited, review of syllabi, exams and other materials were also reported. Becker and Watts (1999) report that department chairs are typically the ones that appoint peer-evaluators. For the most part, this consists of classroom visits, but may also include reviews of syllabi, tests, and other materials. Ironically, White (1995) reports a general reluctance for classroom visits, which are typically only required for junior faculty.

While there is a wealth of research indicating the validity and reliability of student evaluations of teaching, the same is not true for peer review. Not surprisingly, most of the work has focused on classroom observation (Cohen and McKeachie, 1980; Feldman, 1989b). In general, those studies find a notable lack of reliability. However, the reliability of peer review using teaching portfolios appears to show more promise. Root (1987) finds reliability rates of 0.90 using a common faculty committee to assess research, service and teaching based on portfolios assembled by the individual faculty member. Unfortunately such results are not universal (Centra, 1994). Two key factors that account for differences in reliability rates are the selection and training of peer reviewers. For example, Centra (2000) finds that acceptable reliability rates can be obtained when evaluators are not selected by the individual being reviewed. In general, he suggests that small committees, formed for three-year periods, be used to conduct peer reviews. Others argue that reliability rates also increase when the content of portfolios is relatively uniform (Seldin, 1993; Chism, 1999).

Partly because of reliability problems and partly because of faculty antipathy, the focus appears to be turning from classroom observation towards committee review of teaching portfolios. This begs the question of "what to review?" Most argue that portfolios should include a broad range of sample work, syllabi, test, etc. There is universal agreement that peers should receive training that includes methods, standards, and criteria. At a minimum, small committees of between three to six reviewers should be used. Seldin (1999) recommends the following mandatory elements: reflective instructor statement about approach, three years of student evaluations, three years of syllabi for all courses taught, innovative instructional material, and evidence of activities to improve one's teaching.[4]

USE OF ASSESSMENTS

Whether by students or peers, the evaluation of teaching can be used for formative or summative purposes. For evaluations to be used formatively, faculty must value the input of the evaluator and be willing to change behavior.

While theoretically student evaluations can be used for formative purposes, the reality is that they are not. Evidence shows that instructors do not use student evaluations to significantly change their course or their classroom behavior. In a study of 195 teachers over a 13-year period, Marsh and Hocevar (1991) found little change in instructor ratings over time. Ironically, Roche and Marsh (2002) found that student evaluations did change instructor perceptions about their own teaching, as their self-assessments converged with students over time. Thus, student assessments are used by instructors to provide information about the quality of their teaching, but they do not motivate changes in teaching practices. This is precisely why many scholars argue for the importance of alternative systems to complement the use of student evaluations.

Unlike student evaluations, peer evaluation is (theoretically) more likely to result in changes in behavior because the evaluator is qualified to judge the most important aspects of good teaching (Becker, 2000). Because of this, many colleges and universities have established "teaching centers", which, along with supporting scholarship on teaching and learning, typically offer consultation services. Since these centers typically do consultations with faculty directly, such feedback is formative in nature. There is considerable evidence that student evaluations supplemented by professional peer consultations can lead to significant improvements in teaching (Penny and Coe, 2004). One must be cautious in making broad generalizations about the effectiveness of this process as faculty typically self-select into consultations. Also, these groups are not disciplinary experts and thus are limited in the areas of expertise.

The desire to supplement student evaluations helps explain the growing interest in peer evaluation for formative purposes. At the same time, there is little evidence that peer evaluation plays much of a role in summative assessments. Becker and Watts (1999) report that peer assessments were a relatively small part of the overall assessment of teaching used in determining annual merit raises (24, 18, 14, and 11 percent of the teaching assessment at Bachelor, Master's, Doctoral and Research institutions, respectively). Since peer evaluation in economics departments does not appear to be either widespread or highly valued, there is little reason to believe that faculty members have the incentive to respond to potential criticism from their disciplinary peers.

While there is little evidence that faculty use student evaluations for formative purposes, these evaluations do appear to play an important role in summative decisions. In the absence of widespread peer review, any summative assessment of teaching relies predominantly on student evaluations. Moreover, there is evidence that teaching performance is factored into the determination of annual raises, tenure and promotion decisions (Becker and Watts, 1999). Not surprisingly, the relative importance of teaching differs by type of institution. For annual raises, teaching is important for 43 percent of Bachelor, 38 percent of Master's, 37 percent of Doctoral, and 27 percent of Research institutions. The rates are similar for tenure and promotion decisions.

Overall, the evidence suggests that (1) unless used in conjunction with consultation services, faculty do not use student evaluations for formative purposes; (2) peer review in

economics departments is not widely used and, when it is, it is not highly valued; (3) summative decisions are being made almost exclusively on the basis of student end-of-course evaluations; and (4) evaluations of teaching play a significant role in decisions regarding annual merit raises, tenure, and promotions.

CONCLUSION

Despite the plea over a decade ago by Becker and Watts (1999) for economics departments to invest time and energy in using peer review of teaching, it appears little has changed. While teaching effectiveness is a major factor in determining raises and promotions at nearly every institution, economics departments appear to be content to leave such decisions in the hands of undergraduate students. While the research shows students can potentially assess some of the characteristics of good teaching, there is strong reason to believe that biases exist in student evaluations. Even if student evaluations are reliable, they cannot validly assess many of the factors that relate to learning. This is the best argument for peer review. Only colleagues have the disciplinary expertise to assess whether course material is appropriate, whether it facilitates learning and academic challenge and whether it meets departmental goals.

This chapter highlights the need for economists to do more rigorous research on the relative validity of different methods of assessing teaching effectiveness. With few exceptions, the literature focuses almost exclusively on student evaluations. Though a trend towards the increased use of peer review appears to be taking place, in economics and across the academy, there has been little research to validate peer review. Moreover, most research on the effectiveness of teaching has been conducted in other disciplines or at institutional levels. But, as Becker and Watts (1999) argue, there is reason to believe that teaching in economics is different, with a unique approach and specific learning outcomes. As the recent evidence from Weinberg et al. (2009) shows, this may even extend to differences across courses. Taken together, considerably more research into the assessment of teaching *in economics* needs to be undertaken.

NOTES

1. There are other methods that are more formative in nature, such as mid-semester evaluations. In addition, many institutions now offer teaching consultations to faculty.
2. There are a number of excellent meta-analyses on the topic. See Centra (1993), Paulsen (2002), and Marsh (2007).
3. For more information about good research design, see Chapter 35, "Data Resources and Econometric Techniques", in this volume.
4. See Chism (1999) for more about how to design and operate peer evaluations systems.

REFERENCES

Arreola, R. (2000), *Developing a Comprehensive Faculty Evaluation System: A Handbook for College Faculty and Administrators on Designing and Operating a Comprehensive Faculty Evaluation System,* Bolton, US: Anker.

Avery, R. J., W.K. Bryan, A. Mathios, H. Kang and D. Bell (2006), "Electronic course evaluations: Does an online delivery system influence student evaluations?", *Journal of Economic Education*, **37** (1), 21–37.
Becker, W.E. (2000), "Teaching economics in the 21st Century", *The Journal of Economic Perspectives,* **14** (1), 109–19.
Becker, W.E. and M. Watts (1999), "The state of economic education: How departments of economics evaluate teaching", *The American Economic Review*, **89** (2), 334–49.
Cashin, W. E. (1999), "Student ratings of teaching: Uses and misuses", in Peter Seldin (ed.), *Current Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions,* Bolton, US: Anker, pp. 24–44.
Centra, J. A. (1993), *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness,* San Francisco, US: Jossey-Bass.
Centra, J. A. (1994), "The use of the teaching portfolio and student evaluations for summative evaluation", *Journal of Higher Education,* **65** (5), 555–70.
Centra, J. A. (2000), "Evaluating the teaching portfolio: A role for colleagues", *New Directions for Teaching and Learning*, **83** (Fall), 87–93.
Chism, N. (1999), *Peer Review of Teaching: A Sourcebook,* Bolton, US: Anker.
Cohen, P. A. (1987), "A Critical Analysis and Reanalysis of the Multi-section Validity Meta-Analysis", Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20–24).
Cohen, P. A., and W.J. McKeachie (1980), "The role of colleagues in the evaluation of college teaching", *Improving College and University Teaching,* **28** (4), 147–54.
Feldman, K. A. (1989a), "The association between student ratings of specific instructional dimensions and student achievement", *Research in Higher Education,* **30** (6), 583–645.
Feldman, K. A. (1989b), "Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers", *Research in Higher Education,* **30** (2), 113–35.
Hardy, N. (2003), "Online ratings: Fact and fiction", *New Directions for Teaching and Learning*, **96** (Winter), 31–8.
Marsh, H. W. (1987), "Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research", *International Journal of Educational Research,* **11** (3), 253–388.
Marsh, H. W. (2007), "Students' evaluations of university teaching", in R. P. Perry and J. C. Smarts (eds), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-based Perspective*, Netherlands: Springer, pp. 319–83.
Marsh, H. W. and M. J. Dunkin (1997), "Students' evaluations of university teaching: A multi-dimensional perspective", in R. P. Perry and J. C. Smart (eds), *Effective Teaching in Higher Education: Research and Practice*, New York, US: Agathon Press, pp. 241–320.
Marsh, H. W. and D. Hocevar (1991), "Students' evaluations of teaching effectiveness: the stability of mean ratings of the same teachers over a 13-year period", *Teaching and Teacher Education,* **7** (4), 303–14.
Paulsen, Michael B. (2002), "Evaluating teaching performance", *New Directions for Institutional Research*, **114** (Summer), 5–18.
Penny, A. R. and R. Coe (2004), "Effectiveness of consultation on student ratings feedback: A meta analysis", *Review of Educational Research*, **74** (2), 215–53.
Roche, L. A. and H. W. Marsh (2002), "Teaching self-concept in higher education: Reflecting on multiple dimensions of teaching effectiveness," in N. Hativa and P. Goodyear (eds), *Teacher Thinking, Beliefs and Knowledge in Higher Education,* Dordrecht: Kluwer, pp. 179–218.
Root, L. S. (1987), "Faculty evaluation: Reliability of peer assessments of research, teaching and service", *Research in Higher Education*, **26** (1), 71–84.
Seldin, Peter (1993), *Successful Use of Teaching Portfolios,* Bolton, US: Anker.
Seldin, Peter (1999), *Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions*, Bolton, US: Anker, pp. 1–24.
Weinberg, Bruce A., Belton M. Fleisher and Masanori Hashimoto (2009), "Evaluating teaching in higher education", *The Journal of Economic Education*, **40** (3), 227–61.
White, Lawrence J. (1995), "Efforts by departments of economics to assess teaching effectiveness: Results of an informal survey", *The Journal of Economic Education*, **26** (1) 81–85.