# Evaluating Course Evaluations: An Empirical Analysis of a Quasi-Experiment at the Stanford Law School, 2000-2007

Daniel E. Ho and Timothy H. Shapiro

**Abstract**

In Fall 2007, the Stanford Law School implemented what were thought to be innocuous changes to its teaching evaluations. In addition to subtle wording changes, the law school switched the default format for all second- and third-year courses from paper to online. No efforts to anchor the new evaluations to the old were made, as the expectation was that the ratings would be comparable. Amassing a new dataset of 34,328 evaluations of all 267 unique instructors and 350 courses offered at Stanford Law School from 2000-07, we show they are not. This unique case study provides an ideal opportunity to shed empirical light on long-standing questions in legal education about the design, improvement, and implementation of course evaluations. Using finely tuned statistical methods, we demonstrate considerable sensitivity of evaluation responses to wording, timing, and implementation. We offer suggestions on how to maximize information and comparability of evaluations for any institution contemplating reform of its teaching evaluation system.

## Introduction

Prompted by an ailing, twenty year-old scanner, the Stanford Law School implemented seemingly innocuous changes to its teaching evaluations in Fall 2007. Abandoning a law school-specific evaluation that had been in place for over ten years, the school opted for a more extensive, university

**Daniel E. Ho** is Assistant Professor of Law and Robert E. Paradise Faculty Fellow for Excellence in Teaching and Research at Stanford Law School. E-mail: dho@law.stanford.edu; URL: http://dho.stanford.edu.

**Timothy H. Shapiro** is a research fellow at Stanford Law School. E-mail: tshapiro@.stanford.edu.

We thank Faye Deal, George Fisher, Cathy Glaze, Deborah Hensler, Mark Kelman, Larry Kramer, Chidel Onuegbu, Ginny Turner, and Norm Spaulding for helpful comments, and Chidel Onuegbu for help in data collection.

wide evaluation system. Old questions were replaced by subtly altered university-standard questions. Second- and third-year students were asked to respond via an online rating system, in lieu of paper evaluations submitted in class. The primary goal of these changes was to improve the information about teaching quality at the law school.

The desire to optimize the evaluation process is understandable. Teaching evaluations play a crucial role in hiring, promotion, and tenure decisions. Student evaluations of teaching remain a prevalent measure in undergraduate institutions,[1] as well as business,[2] medical,[3] and law schools.[4] Despite widespread use, consensus on their validity remains elusive, with scholars highlighting interpretation difficulties,[5] non-correspondence between evaluations and student performance,[6] and lack of comparability, validity, or reliability.[7] Yet even among

1. See Pamela J. Eckard, Faculty Evaluation: The Basis for Rewards in Higher Education, 57 Peabody J. Educ. 94, 96 (1980); Timothy J. Gallagher, Embracing Student Evaluations of Teaching: A Case Study, 28 Teaching Sociology 140 (2000); and Gordon E. Greenwood and Howard J. Ramagli, Jr., Alternatives to Student Ratings of College Teaching, 51 J. Higher Educ. 673, 674 (1980).

2. See Curt J. Dommeyer, et al., Attitudes of Business Faculty Towards Two Methods of Collecting Teaching Evaluations: Paper v. Online, 27 Assessment & Evaluation in Higher Educ. 455 (2002); and Keith G. Lumsden, Summary of an Analysis of Student Evaluations of Faculty and Courses, 5 J. Econ. Educ. 54 (1973).

3. See Jennifer R. Kogan and Judy A. Shea, Course Evaluation in Medical Education, 23 Teaching and Tchr. Educ. 251 (2007); and Anthony M. Paolo et al., Response Rate Comparisons of E-mail- and Mail-Distributed Student Evaluations, 12 Teaching & Learning in Med. 81 (2000).

4. See William Roth, Student Evaluation of Law Teaching: Observations, Resource Materials, and a Proposed Questionnaire 1 (Ass'n of Am. L. Sch. Washington D.C. 1983); Richard L. Abel, Evaluating Evaluations: How Should Law Schools Judge Teaching, 40 J. Legal Educ. 407, 412 (1990); and Judith D. Fischer, How to Improve Student Ratings in Legal Writing Courses: Views from the Trenches, 34 U. Balt. L. Rev. 199 (2004).

5. See Daniel Gordon, Does Law Teaching Have Meaning? Teaching Effectiveness, Gauging Alumni Competence, and the MacCrate Report, 25 Fordham Urb. L.J. 43 (1998); and Christopher T. Husbands, Variation in Students' Evaluations of Teachers' Lecturing in Different Courses on Which They Lecture: A Study at the London School of Economics and Political Science, 33 Higher Educ. 51 (1997).

6. See William E. Becker and Michael Watts, How Departments of Economics Evaluate Teaching, 89 Am. Econ. Rev. 334 (1999); and Greenwood and Ramagli, Alternatives to Student Ratings, *supra* note 1, at 675.

7. See Charles F. Eiszler, College Students' Evaluations of Teaching and Grade Inflation, 43 Res. Higher Educ. 483 (2002) (discussing correlation between teacher ratings and grade inflation); Laura I. Langbein, The Validity of Student Evaluations of Teaching, 27 PS: Pol. Sci. & Pol. 545, 552 (1994) (examining whether ratings are responsive to teacher popularity); Theodore C. Wagenaar, Student Evaluation of Teaching: Some Cautions and Suggestions, 23 Teaching Sociology 64 (1995) (highlighting grade inflation and popularity as major sources of confounding); and Donald H. Naftulin et al., The Doctor Fox Lecture: A Paradigm of Educational Seduction, 48 J. Med. Educ. 630, 634 (1973) (finding that "students can be effectively 'seduced' into an illusion of having learned if the lecturer simulates a style of authority and wit").

detractors, a majority supports the use of student evaluations as *one* component of instruction assessment.[8] As applied to legal education specifically, the Association of American Law Schools has made the incorporation and interpretation of student evaluations an ongoing priority, soliciting and publishing research on the subject over the past twenty-five years.[9]

Stanford Law School's unique quasi-experiment provides an ideal opportunity to study long-standing questions in legal education about the design, improvement, and implementation of course evaluations. Do students respond measurably to subtle changes in evaluation questions? How easily can we compare responses between different types of evaluations? Does the timing and format of administration matter? And, given unending calls for reform, how can law schools reliably improve teaching assessment? To address these questions, we amass a new dataset of 34,328 evaluations of all 267 unique instructors and 350 courses offered at Stanford Law School from 2000-07. We find that in spite of the laudable goal to gather more information, *less* information may have been gained as a result of these changes. Strong evidence shows that the online system results in fewer respondents, and that the format and wording changes affected those respondents' answers.

Our examination allows us to draw key pragmatic empirical lessons about the evaluation of teaching evaluations in legal education. First, our case study demonstrates the perils of attempting to draw comparisons across different systems, and illustrates the need to carefully anchor old and new evaluations. We document dramatic effects of wording and timing changes on evaluations, well-known in the survey literature.[10] Although superficially similar, subtle

8. See Abel, Evaluating Evaluations, *supra* note 4, at 452; Becker and Watts, How Departments of Economics Evaluate Teaching, *supra* note 6; Sylvia d'Appolonia and Philip C. Abrami, Navigating Student Ratings of Instruction, 52 Am. Psychol. 1198, 1205 (1997); John A. Centra, Research Productivity and Teaching Effectiveness, 18 Res. Higher Educ. 379 (1983); Eiszler, College Students' Evaluations, *supra* note 7, at 499; Gallagher, Embracing Student Evaluations, *supra* note 1, at 142; Langbein, The Validity of Student Evaluations, *supra* note 7, at 552; and Herbert W. Marsh and Lawrence A. Roche, Making Students' Evaluation of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility, 52 Am. Psychol. 1187, 1193 (1997).

9. See, e.g., Roth, Student Evaluation, *supra* note 4; Arthur Best, Student Evaluations of Law Teaching Work Well: Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree 3 (June, 2006) (presentation at AALS conference "New Ideas for Law School Teachers: Teaching Intentionally"); Gerald F. Hess, Heads and Hearts: The Teaching and Learning Environment in Law School, 52 J. Legal Educ. 75, 97 (2002); Robert M. Lloyd, On Consumerism in Legal Education, 45 J. Legal Educ. 551 (1995); Richard S. Markovitz, The Professional Assessment of Legal Academics: On the Shift from Evaluator Judgment to Market Evaluations, 48 J. Legal Educ. 417 (1998); Eric W. Orts, Quality Circles in Law Teaching, 47 J. Legal Educ. 425 (1997); and Victor G. Rosenblum et al., Report of the AALS Special Committee on Tenure and the Tenuring Process, 42 J. Legal Educ. 477, 495 (1992).

10. See Vic Barnett, Sample Survey: Principles and Methods 176 (3d ed., New York, 2002); William A Belson, The Design and Understanding of Survey Questions 240 (Aldershot, Eng., 1981); Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski, The Psychology of Survey Response (Cambridge, 2000); Graham Kalton, Martin Collins and Linsay

differences in question wording may systematically affect evaluations, threatening comparability of evaluations across institutions or time. At Stanford, the new evaluations shifted both the mean and variance on the 5-point rating scale. The inadvertent result can be dramatic when considering a "cutoff" rule: for the same course, an instructor has a 35 percent probability of falling below 4.5 using the old evaluations, but this probability jumps to 59 percent with the new evaluations. As far as we are aware, this study is the first to systematically document such design effects on course evaluations. We suggest that well-known techniques to anchor evaluations be adopted to remedy these incomparabilities.

Second, we demonstrate a primary pitfall to adopting an online response system. While online evaluations have many potential advantages, including increased survey flexibility,[11] shorter response times,[12] and reduced costs,[13] they may suffer from generally high and nonrandom nonresponse[14] and more negative tenor,[15] threatening

Brock, Experiments in Wording Opinion Questions, 27 Applied Stat. 149-161 (1978); Sam G. McFarland, Effects of Question Order on Survey Responses, 45 Pub. Opinion Q. 208 (1981); and Mary K. Serdula et al., Effects of Question Order on Estimates of the Prevalence of Attempted Weight Loss, 142 Am. J. Epidemiology 64 (1995).

11. See Mick P. Couper, Web Surveys: A Review of Issues and Approaches, 64 Pub. Opinion Q. 464, 465 (2000) (discussing how web surveys simplify the delivery of multimedia content); and Benjamin H. Layne, Joseph R. DeCristoforo, and Dixie McGinty, Electronic Versus Traditional Student Ratings of Instruction, 40 Res. Higher Educ. 221, 229 (1999) (finding that students completing electronic evaluations were more likely to provide written comments).

12. See Kim B. Sheehan and Sally J. McMillan, Response Variation in E-Mail Surveys: An Exploration, 39 J. Advertising Res., July-Aug. 1999, at 45, 46 (1999) (conducting meta-analysis and finding shorter response time for e-mail compared with conventional mail).

13. See Cihan Cobanoglu, Bill Warde, and Patrick J. Moreo, A Comparison of Mail, Fax, and Web-Based Survey Methods, 43 Int'l J. Market Res. 441, 449 (2001); Couper, Web Surveys, *supra* note 11, at 464; and Michael D. Kaplowitz, Timothy D. Hadlock, and Ralph Levine, A Comparison of Web and Mail Survey Response Rates, 68 Pub. Opinion Q. 94, 98 (2004).

14. See Mick P. Couper, Johnny Blair, and Timothy Triplett, A Comparison of Mail and E-Mail for Survey of Employees in Federal Statistical Agencies, 15 J. Official Stat. 39, 46 (1999); and Sheehan and McMillan, Response Variation in E-Mail Surveys, *supra* note 12, at 46. Regarding online teaching evaluations in particular, see Curt J. Dommeyer et al., Gathering Faculty Teaching Evaluations By In-Class and Online Surveys: Their Effects on Response Rates and Evaluations, 29 Assessment & Evaluation in Higher Educ. 611, 618 (2004); and Layne et al., Electronic versus Traditional Student Ratings, *supra* note 11, at 226 (60 percent response in-class vs. 48.7 percent online).

15. See Miguel A. Vallejo et al., Psychological Assessment via the Internet: A Reliability and Validity Study of Online (vs. Paper-and-Pencil) Versions of the General Health Questionnaire-28 (GHQ-28) and the Symptoms Check-List-90-Revised (SCL-90-R), 9 J. Med. Internet Res. 2 (2007) (finding that "paper and pencil scores were higher than online ones" for standardized health questionnaires, and that a large bank of data will be necessary to anchor online responses to normative questions in the future). See also Mei Alonzo and Milam Aiken, Flaming in Electronic Communication, 36 Decision Support Sys. 205 (2004); Sara Keisler, Jane Siegel, and Timothy W. McGuire, Social Psychological Aspects

the validity of any summary results.[16] Our analysis confirms such bias in the law school context. We highlight specific factors in the law school's implementation that exacerbated these pitfalls, and provide suggestions for how to maximize response and comparability.

More generally, this study contributes to the empirical understanding of the dynamics of the student evaluation process.[17] Our findings inform sophisticated use of teaching evaluations, as well as the large body of literature that uses instructor evaluations to study aspects of legal education, such as gender[18] and minority discrimination,[19] and the relationship between teaching and scholarship.[20] Our examination reveals considerable temporal trends in survey responses, suggesting nonrandom nonresponse bias across terms and instructors, and a strong upward trend in mean evaluations.[21] We demonstrate how to account for such trends to consistently and effectively learn from evaluations. One of the substantial collateral benefits of our analysis is that we gain considerable insight into the transformation and development of the law school over the past eight years, as we show below.

We proceed as follows. First we document how the evaluation system was reformed in Fall 2007. Then we describe the empirical scaling problem of how to equate the new evaluations with the old system. The next section describes the dataset we use to shed light on what impact the Fall 2007 evaluations may have had, after which we present results of our analysis, which brings to bear finely

of Computer-Mediated Communication, 39 Am. Psychologist 1123, 1129 (1984); and John Suler, The Online Disinhibition Effect, 7 CyberPsychol. & Behav. 321 (2004).

16. See Stephen J. Sills and Chunyan Song, Innovations in Survey Research: An Application of Web-Based Surveys, 20 Soc. Sci. Computer Rev. 22, 26 (2002); and Bruce Ravelli, Anonymous Online Teaching Assessments: Preliminary Findings 7 (June 14, 2000) (unpublished manuscript, at www.eric.ed.gov, #ED445069) ("[S]tudents expressed the belief that if they were content with their teacher's performance, there was no reason to complete the survey.").

17. See Eiszler, College Students' Evaluations, *supra* note 7 (documenting evaluations' relationship to grade inflation); Christopher J. Fries and R. James McNich, Signed Versus Unsigned Student Evaluations of Teaching: A Comparison, 31 Teaching Soc. 333, 341 (2003) (documenting effects of anonymity in course evaluations); and Stephen Shmanske, On the Measurement of Teacher Effectiveness, 19 J. Econ. Educ. 307 (1988) (documenting evaluations' relationship to student performance).

18. See, e.g., Kenneth A. Feldman, College Students' Views on Male and Female College Teachers: Part II–Evidence from Students' Evaluations of Their Classroom Teachers, 34 Res. Higher Educ. 151 (1993); Joan M. Krauskopf, Touching the Elephant: Perceptions of Gender Issues in Nine Law Schools, 44 J. Legal Educ. 311, 312 (1994); Langbein, The Validity of Student Evaluations, *supra* note 7, at 552; and Deborah L. Rhode, Taking Stock: Women of All Colors in Legal Education, 53 J. Legal Educ. 475, 480 (2003).

19. See, e.g., Abel, Evaluating Evaluations, *supra* note 4, at 407.

20. See, e.g., Centra, Research Productivity, *supra* note 8; and James Lindgren and Allison Nagelberg, Are Scholars Better Teachers? 73 Chi.-Kent L. Rev. 823 (1998).

21. See Robert M. Groves et al., Survey Methodology 167 (New York, 2004); Robert M. Groves et al., Survey Nonresponse (New York, 2002); Sharon L. Lohr, Sampling: Design and Analysis 225 (Pacific Grove, Cal., 1999).

tuned statistical methods (subclassification, matching, multilevel modeling, and nonparametric bounds) to address potential confounding factors. We conclude with concrete implications on the use, evaluation, and reform of teaching evaluations.

**The Change in Evaluations**

Prior to Fall 2007, Stanford Law School employed its own evaluation system, using seven questions with discrete responses (e.g., do you agree that "readings were appropriate and useful"), and four "write-in" questions (e.g., "what improvements, if any, would you suggest"). To increase the information about teaching quality, the law school adopted a university wide questionnaire in Fall 2007, which contained *different questions* and increased the number of questions with discrete responses to eighteen and the number of write-in questions to six.

Table 1 reproduces the primary question of interest for the old and new evaluation system. "Overall effectiveness" has conventionally served as the main summary of the evaluations, so we focus on it for the remainder of this article.[22] The old question about overall effectiveness, presented in the left column, asked students to respond to whether "overall[] the instructor was effective as a teacher." Responses ranged from "disagree strongly" to "agree strongly," and were conventionally tabulated from 1 to 5, with 5 representing the most positive rating of "agree strongly." We follow this convention in reporting our results. Note that the old response scale deviated from conventional (Likert[23]) scales in subtle ways.[24] A conventional scale might run from "strongly disagree," "disagree," "neither agree nor disagree," "agree," to "strongly agree." Likely, the addition of "somewhat" artificially increased mean ratings on the old system.

The right column presents the new question, which asked students to assess "the instructor's overall teaching." Responses range from "poor," "fair," "good," "very good," to "excellent." These responses were again transformed to a numerical 1-5 scale, with 5 representing "excellent."[25]

22. The focus on a single question of overall teaching effectiveness appears common across institutions. See, e.g., Roth, Student Evaluation, *supra* note 4, at II-3 ("for administrative purposes, the most useful item is a composite (or 'global') question pertaining to overall teaching quality"); and Gallagher, Embracing Student Evaluations, *supra* note 1, at 142 ("Departments...tend to give greater weight to the global items in assessing teaching quality").

23. See generally Rensis Likert, A Technique for the Measurement of Attitudes, 140 Archives Psychol. 55 (1932).

24. Roth, Student Evaluation, *supra* note 4, at B-8 (reproducing seventy law schools' teaching evaluation forms, forty-two of which use the "poor" to "excellent" scale, thirteen of which use the "disagree" to "agree" scale, and two of which use the response categories "somewhat agree" and "somewhat disagree").

25. For empirical studies on the effects of response categories, see Alan J. Klokars and Midori Yamagishi, The Influence of Labels and Positions in Rating Scales, 25 J. Educ. Measurement

**Table 1**

| **Old Evaluation** | **New Evaluation** |
|---|---|
| "Overall: The Instructor was effective as a teacher | "The Instructor's overall teaching" |
| **Responses:** | **Responses:** |
| 1. Disagree Strongly | 1. Poor |
| 2. Disagree Somewhat | 2. Fair |
| 3. Neutral | 3. Good |
| 4. Agree Somewhat | 4. Very Good |
| 5. Agree Strongly | 5. Excellent |

Table 1: Primary questions of interest in course evaluations before Fall 2007 on the left and in Fall 2007 on the right. Conventionally, a 1-5 scale was assigned to each answer, with the mean reported publicly as a single numerical summary of course evaluations.

In addition to changing the questions, the default format for all second- and third-year courses was changed from a paper evaluation, typically administered on the last day of the semester, to an online evaluation that could be submitted over a long time window, from before class ended to (potentially) after the final examination. To improve the response rate, most departments at Stanford withhold or delay grades from undergraduates who have not submitted course evaluations.[26] Due to calendar differences and the independence of the law school registrar, the law school did not have the same capacity to withhold grades.

Both the old and new evaluations were transformed to the seemingly same 1-5 scale. Initially, it was believed that the numerical scales would be equivalent, so no efforts to anchor the new evaluations were made. After all, they bear superficial resemblance, post-transformation.

After cursory examination of evaluation responses in Fall 2007—yielding some perplexing results—the dean of the law school asked us to examine more systematically whether the new evaluations may have inadvertently affected responses. Our analysis reveals that the use of the scales is unlikely to be equivalent.[27]

85 (1988); Tony C.M. Lam and Alan J. Klokars, Anchor Point Effects on the Equivalence of Questionnaire Items, 19 J. Educ. Measurement 317 (1982); and Tony C.M. Lam and Joseph J. Stevens, Effects of Content Polarization, Item Wording, and Rating Scale Width on Rating Response, 7 Applied Measurement in Educ. 141 (1994).

26. On incentivizing response, see Roth, Student Evaluation, *supra* note 4, at II-2; M. Berlin et al., An Experiment in Monetary Incentives, *in* Proceeding of the Survey Research Methods Section of the American Statistical. Association 393 (Alexandria, Va., 1992); and Dommeyer, Gathering Faculty Teaching Evaluations, *supra* note 14, at 619 (suggesting that early grade feedback significantly increases response rate compared with a control group).

27. See Belson, The Design and Understanding of Survey Questions, *supra* note 10; and

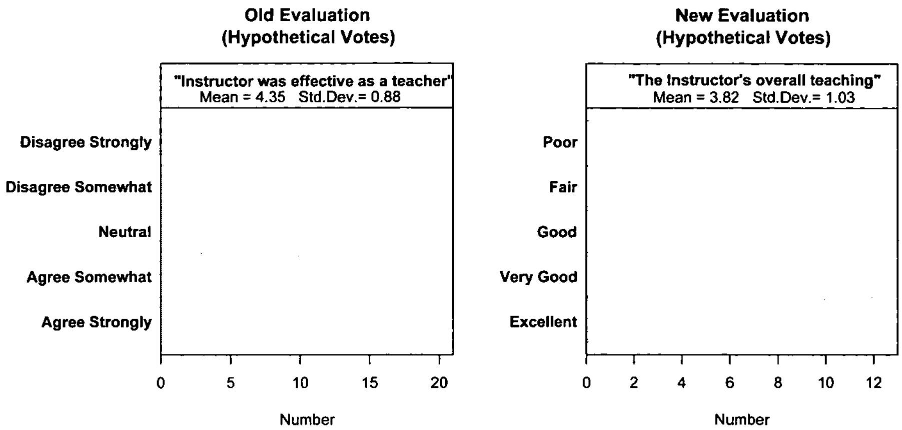


Figure 1: Hypothetical votes to illustrate empirical scaling problem. While the mean difference is considerable, it remains unclear whether the courses are in fact distinguishable in student perception due to the different questions.

## The Empirical Scaling Problem

There are strong reasons to doubt the comparability of raw quantitative averages for two different questions. Not only do the numbers represent qualitatively different responses, but the new evaluation system may affect both the mean *and the variance* (or the entire distribution) of the response scale. As a result, any simple mean adjustment (e.g., adding 0.3 to the new evaluations) may equally mislead. In the literature on test equating (e.g., scaling the SAT so that it remains comparable across different test takers and exams),[28] typical solutions are to equate tests by common subjects[29] or questions.[30] In our case study, neither is available to anchor our scales due to (1) anonymous evaluations and (2) no overlap of questions on any single form. Randomizing the questions to students in the same course would

Norman Schwartz et al., Rating Scales: Numeric Values May Change the Meaning of Scale Labels, 55 Pub. Opinion Q. 570 (1991).

28. See William H. Angoff, Summary and Derivation of Equating Methods Used at ETS, *in* Test Equating 55 (Paul W. Holland and Donald B. Rubin eds., New York, 1982).

29. See Michael J. Kolen and Robert L. Brennan, Test Equating, Scaling, and Linking 15 (2nd ed., New York, 2004); and Carl N. Morris, On the Foundations of Test Equating, *in* Test Equating 171 (Paul W. Holland and Donald B. Rubin eds., New York, 1982). For a creative approach to creating scale linkages to remedy cultural incomparability, see Gary King, et al., Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research, 98 Am. Pol. Sci. Rev. 191 (2004).

30. See Kolen and Brennan, Test Equating, *supra* note 29, at 103; Howard Wainer, et al., Computerized Adaptive Testing: A Primer 12 (Hillsdale, N.J., 1990); and Nancy S. Petersen, Michael J. Kolen, and H.D. Hoover, et al., Scaling, Norming, and Equating, *in* Educational Measurement 221 (Robert B. Linn ed., 3d ed., New York, 1989).

have alternatively solved the scaling problem, but was not attempted.[31] The empirical scaling problem hence becomes daunting, and the new evaluation system poses at least two distinct problems, in substance and form.

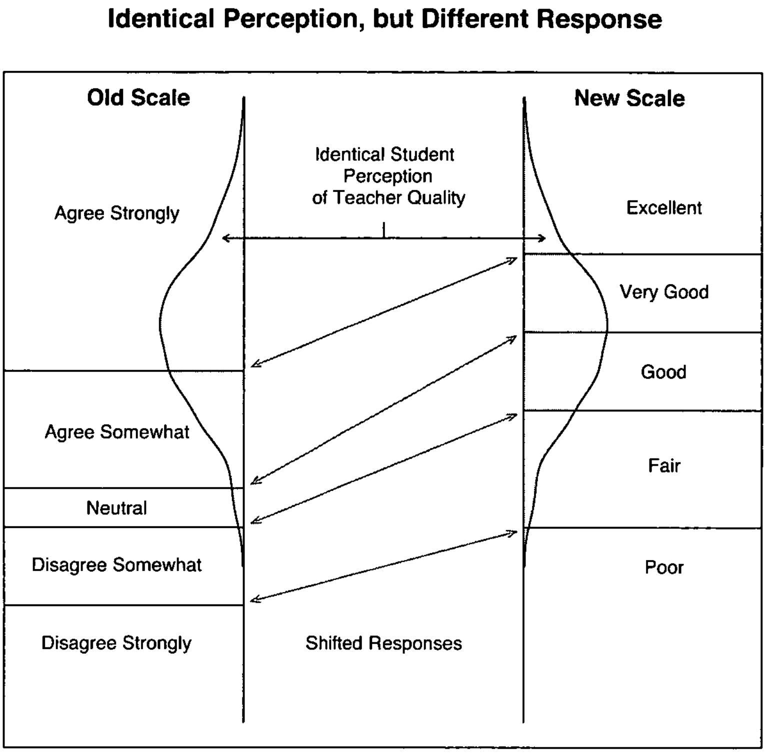


Figure 2: The gray-shaded densities are mirror images and represent the latent "true" perception of a course consistent with observed ratings in Figure 1. New scales reposition the cutpoints that translate latent perception into observable responses. This figure shows that identical perceptions may result in dramatically different numerical responses on course evaluations.

**Substance: Different Questions.** In substance, the questions are different, albeit in subtle ways. The old evaluation asked students to respond to the assertion that "overall[] the instructor was effective as a teacher," while the new evaluation asked students to evaluate "the instructor's overall teaching," with no default assertion or reference to effectiveness. Intuitively, "neutral" does not equal "good": a "neutral" response to the former likely represents a worse evaluation than a response of "good" to the latter, but both are represented

31. See generally Donald B. Rubin, Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, 66 J. Educ. Psychol. 688 (1974).

as 3's numerically. In addition, the old responses have a clear "center" of "neutral," with symmetric responses deviating from that center, while the new scale doesn't exhibit that facial symmetry. Figure 1 illustrates this scale shift with histograms of hypothetical votes on the old system in the left panel and the new system in the right panel. The raw distributions differ considerably, with means of 4.35 and 3.82 on the old and new system, respectively. Yet without some way to anchor these scales, responses from one can't be compared to the other. For example, a 3.82 on the new system may even represent a *better* overall evaluation than a 4.32 on the old system.

To formalize the problem, Figure 2 plots distributions of the underlying (latent) perceptions from which the ratings in Figure 1 might be generated. Both left and right gray-shaded densities (smoothened histograms) are mirror images, meaning that student perceptions are identical for the course. Using the old scale, the cutpoints translating the latent perceptions into numerical responses (represented by the areas between the horizontal lines) are quite uneven. The modal category, represented by the large gray mass on top, is "agree strongly," and the second most-used category is "agree somewhat." The new response scale repositions the cutpoints, resulting in more even use of the response categories.[32] In short, in this hypothetical example, different numerical responses are driven solely by student interpretation of the new scale, not any change in the course's perception. Even when the underlying perception is *identical* and when the survey is administered under the same conditions, we might expect different numerical responses.

**Format: Timing, and Paper versus Online.** All conditions, however, were not the same. Compounding the scaling problem was the simultaneous change in evaluation format. With the exception of first-year courses, the new evaluations were submitted online through Stanford's central university website (called "Axess"). Students received an e-mail with a hyperlink to submit course evaluations, and logged onto Axess using student-specific IDs to prevent double-voting.[33] In Fall 2007, the submission window was from December 3-16, from the first day of the university's "End of Quarter Period" (but not the law school's) to the Sunday after university (but not law school) final exams. Because the university still maintains a different calendar but centrally administers the evaluations, students could theoretically submit evaluations before law school courses ended until *after they finished with final exams*. Since the process was centralized, instructors did not necessarily provide time during the last class to submit evaluations, as was customary under the old system.

32. From the perspective of gathering maximum information, such a scale may be more useful, as it is likely to discriminate more between courses and instructors.

33. See Fries and McNich, Signed versus Unsigned Student Evaluations, *supra* note 17 (finding changes in evaluation response under perception of anonymity); and Layne, Electronic Versus Traditional Student Ratings, *supra* note 11, at 229 ("most students felt that the traditional method would have afforded them a higher degree of anonymity than the electronic method, particularly since it had been necessary for them to use their student identification numbers to log in to the system").

In addition to the timing changes, the implementation switch, from paper to online evaluations, may induce different responses (e.g., students who are frustrated by their exam performance may voice grievances online).[34]

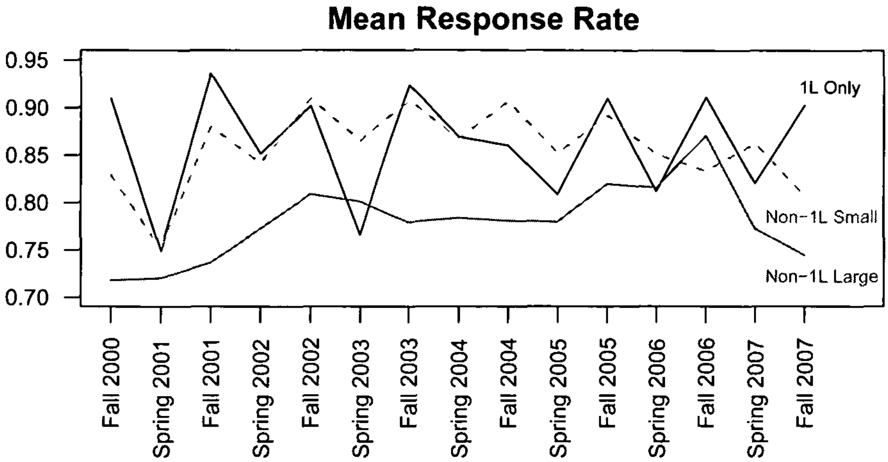


Figure 3: This figure plots the response rate on the *y*-axis against terms on the *x*-axis. The black line represents first-year courses. The solid and dashed gray lines represent upper division courses, for large and small courses, respectively, using the term-specific median enrollment as the cutoff. The response rate for first-year evaluations is as expected in Fall 2007, but it drops for upper division courses.

The new format may also have changed the type and number of students from whom information was gathered. Previous research and our own evidence suggest, for example, that online systems may reduce the response rate.[35] Figure 3 plots the response rate from 2000 to 2007. The black line represents first-year courses, which exhibit a strong fall-spring effect. Since first-year evaluations were administered as usual on paper during the last class, the response rate in Fall 2007 appears as we might expect (upwards in the fall). The dashed and solid gray lines represent small and large upper-division courses (divided by the term-specific median enrollment), respectively. Both drop in Fall 2007 beyond what one might expect. The response rate for large courses appears to drop one term prior to Fall 2007 as well, which was part of the impetus for the shift to online evaluations. While the spirit of gathering more information is to be applauded, the response rate changes may mean that increasingly non-random samples of students are responding to course evaluations.

34. See *supra* note 15.

35. See Couper, Web Surveys, *supra* note 11, at 464; Kaplowitz, Hadlock, and Levine, A Comparison of Web and Mail Survey Response Rates, *supra* note 13, at 94; and Sills and Song, Innovations in Survey Research, *supra* note 16, at 22.

In sum, although not necessarily apparent at first blush, there are strong reasons to expect scale incomparability due to the substantive question change of evaluations,[36] and the shift in the timing[37] and format.[38] Because all of these changes were implemented concurrently (and simultaneous to changes in instructors, courses, and students), empirically estimating the scale effect poses formidable challenges. We turn to these now.

| | Fall 2000 - Spring 2007 | | | Fall 2007 | | |
|---|---|---|---|---|---|---|
| | Mean | SE | N | Mean | SE | N |
| **Course-Level Statistics** | | | | | | |
| Mean Rating | 4.51 | 0.50 | 1219 | 4.34 | 0.48 | 112 |
| Enrollment | 25.60 | 22.67 | 1238 | 23.52 | 17.82 | 112 |
| Response Rate | 0.83 | 0.18 | 1227 | 0.80 | 0.14 | 112 |
| **Semester-Level Statistics** | | | | | | |
| Unique Courses | 67.64 | 13.80 | 14 | 80.00 | | 1 |
| Unique Instructors | 68.93 | 14.10 | 14 | 90.00 | | 1 |

Table 2: Summary statistics for course evaluation dataset. SE indicates standard error, and N indicates the number of observations.

## Data

Our data originates from the law school's Office of Student Affairs and Stanford's Axess system.[39] Appendix A provides details on compilation of the dataset. After cleaning and recoding for consistency, the data encompasses 34,328 evaluations, 267 unique instructors (including full-time faculty, lecturers, and other affiliates of the law school), and 350 unique courses. Table 2 provides basic summary statistics of the dataset for Fall 2000 to Spring 2007 in the left columns and for Fall 2007 in the right columns. Prior to the new system, the mean effectiveness was 4.51, compared to 4.34 on the new system. This raw

36. See Barnett, Sample Survey, supra note 10; Belson, The Design and Understanding of Survey Questions, *supra* note 10; Tourangeau et al., The Psychology of Survey Responses, *supra* note 10; and Schwartz, Rating Scales, supra note 27.

37. See Dommeyer, Gathering Faculty Teaching Evaluations, *supra* note 14, at 618; and Sheehan and McMillan, Response Variation in E-mail Surveys, *supra* note 12, at 46.

38. See Couper, Web Surveys, *supra* note 11, at 465; Kaplowitz et al., A Comparison of Web and Mail Survey Response Rates, *supra* note 13, at 94; Layne et al., Electronic Versus Traditional Student Ratings of Instruction, *supra* note 11, at 229; and Sills and Song, Innovations in Survey Research, *supra* note 16, at 22.

39. The data includes information on courses and instructors for which student evaluations were solicited; while this is appropriate for assessing the impact of the new evaluations, one should be cautious about inferring too much about broader trends as the data was not validated against registrar and employment records.

difference, however, does not likely represent the effect of the new system because course evaluations have been steadily improving over the observation period as class sizes have steadily decreased.

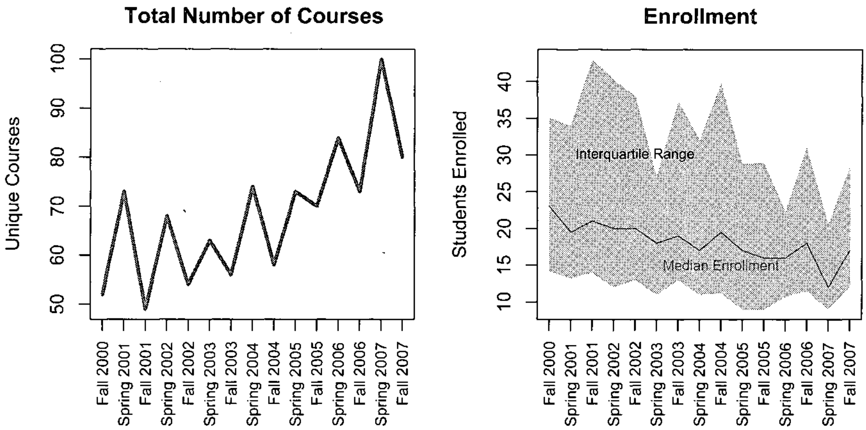


Figure 4: The left panel of this figure plots the number of unique courses offered in a particular semester at Stanford Law School. The right panel shows the decreasing trend in class size, observed in both the median and interquartile range of enrolment.

The left panel of Figure 4 presents the number of unique courses offered over terms. The data shows a considerable fall-spring effect, with roughly ten to twenty additional courses offered in the spring. The number of courses offered each term has consistently increased from 2000 to 2007, from a low of around 50 courses in Fall 2000 and 2001, to 100 courses offered in Spring 2007. The right panel shows a concordant decrease in class sizes. The black line plots the median enrollment, ranging from a high of 23 in Fall 2000 to a low of 12 in Spring 2007. The gray bands present the interquartile range (25th to 75th percentile of enrollment) showing a similar decrease over time and the fall-spring effect. One of the consistent findings in the literature on course evaluations is that evaluations tend to be better for smaller classes.[40] Accounting for this trend therefore is crucial to assessing the effect of the new evaluations.

40. See L. F. Jameson Boex, Attributes of Effective Economics Instructors, 31 J. Econ. Educ. 211, 219 (2000); Albert L. Danielsen and Rudolph A. White, Some Evidence on the Variables Associated with Student Evaluations of Teachers, 7 J. Econ. Educ. 117, 118 (1976); Herbert W. Marsh, Jesse U. Overall, and Stephen P. Kesler, Class Size, Students' Evaluations, and Instructional Effectiveness, 16 Am. Educ. Res. J. 57, 61 (1979); and Kenneth Wood, Arnold S. Linsky, and Murray A. Straus, Class Size and Student Evaluation of Faculty, 45 J. Higher Educ. 524, 528 (1974).

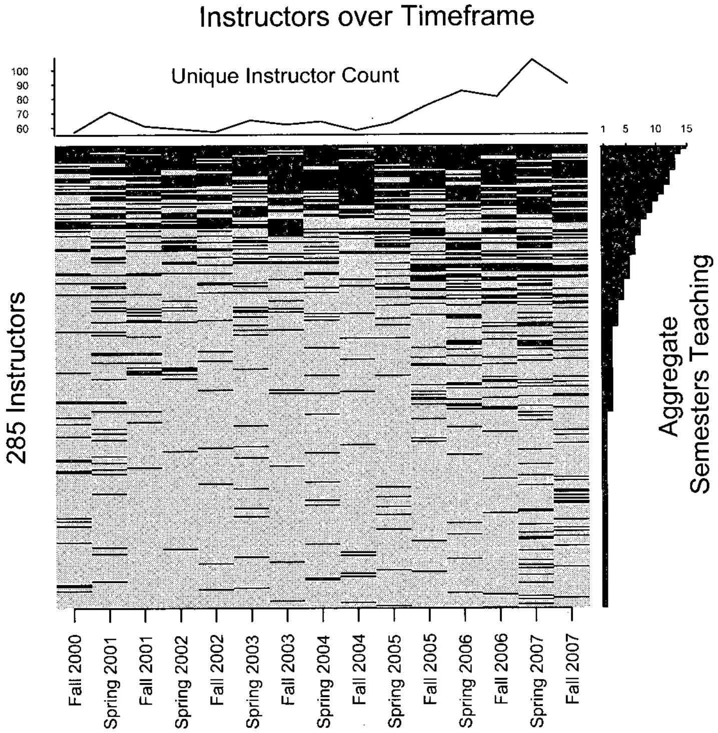


Figure 5: Each row represents a unique instructor from our dataset. Rows are sorted by total courses taught, and randomly ordered within ties. Black cells indicate that a given instructor taught at least one class during a given semester. This figure shows that more instructors were teaching in recent years. While the majority of instructors who have ever taught at the law school teach for only one or two semesters, this figure also shows that 29 percent of the instructors in the dataset teach for five or more semesters. The panel on the right represents the count of semesters taught per instructor. The panel at the top represents the count of unique instructors by semester. Both the number of instructors and the number of courses have increased in recent years.

Figure 5 provides an overview of the continuity of instruction at the law school. Each row represents a unique instructor who taught at least once during the observation period. Cells are black if the instructor taught during a particular term and faint gray otherwise. Two trends emerge from this figure. First, the number of instructors has increased since Fall 2004, as can be seen in the top panel and the increase in the black cells in the right columns on the main panel. These changes pose difficulties for cleanly identifying the effect of the new evaluations, as ratings changes may be due to changes in the faculty. Second, as summarized by the right histogram, while most instructors have taught only one or two courses at the law school, almost one third of instructors have taught for five or more semesters. The multilevel model we outline below exploits the continuity of instructors and courses to account for changes in who teaches and what is taught at the law school.

## Statistical Analysis

### *Raw Differences*

To obtain an initial estimate of the impact of the new evaluations, we examine the distribution of ratings over time. Figure 6 plots raw numerical means of the primary outcome measure on the $y$-axis against terms on the $x$-axis. Each circle represents the mean rating for a course in a given term, with the area proportional to course enrollment. This figure shows two principal patterns. First, course evaluations have improved markedly over time, as can be seen by the increase in the horizontal bars (representing means) over time. Such improvements may be attributable to the drop in class sizes and increase in course offerings, as other factors, such as the first-year core classes and overall admitted class size, have stayed roughly constant over this time period. Second, focusing on the distribution on the right, we detect a marked shift in the evaluations in Fall 2007. There is much more "mass," as indicated by dark overlaps of circles, below 4, and the horizontal bar, representing the grand mean, drops considerably. This figure, depicting all the data, provides strong suggestive evidence that the new evaluation system mattered.

To assess how robust this raw difference is, we pursue several strategies below. At the outset, we caution that because (a) we observe only one term with the new evaluation system, and (b) many other dimensions of the law school are simultaneously changing, these findings are necessarily limited. If something unique happened in Fall 2007 (e.g., general school-wide malaise), we will falsely attribute any scale shift to the evaluation system.

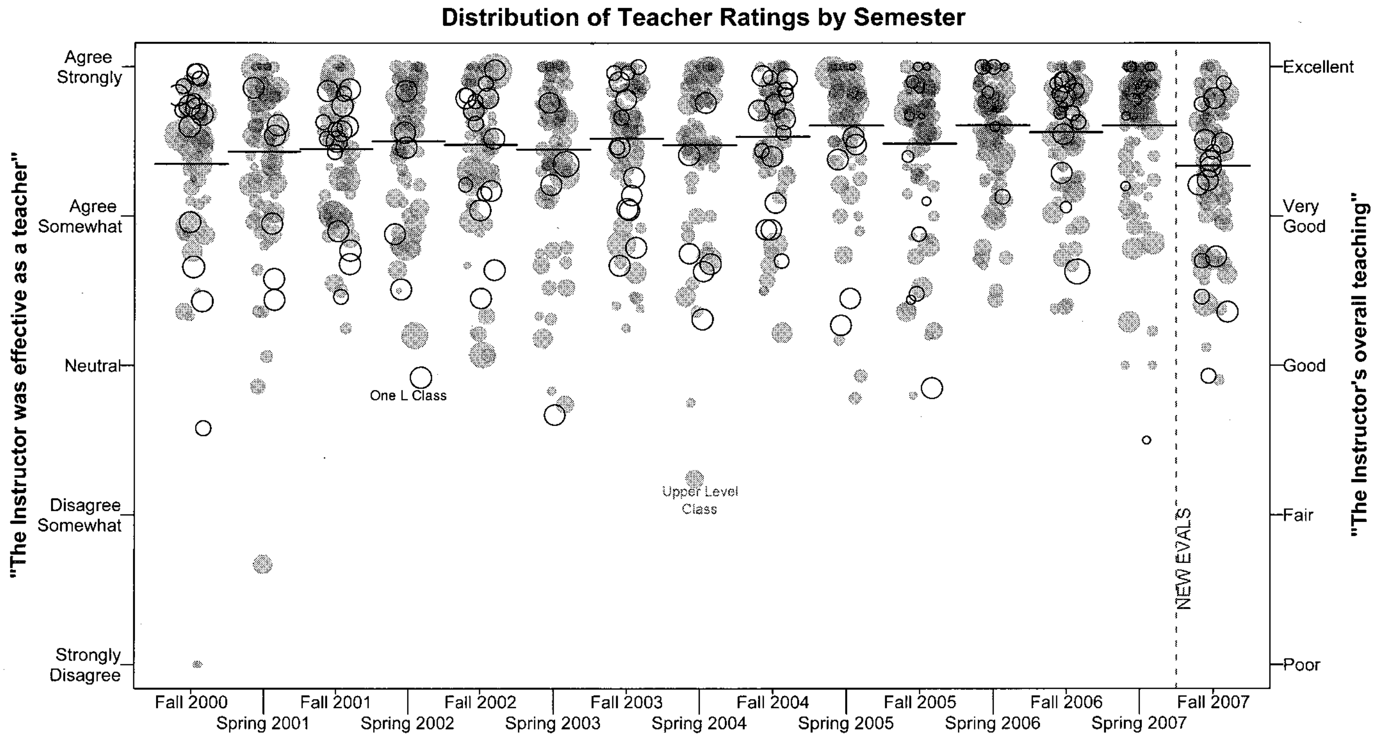


Figure 6: This figure plots the primary outcome data of mean evaluations over terms. Each circle represents the mean rating for a course, with the area proportional to enrollment. Circles are randomly jittered horizontally for visibility. First-year courses are hollow black, and upper division courses are shaded gray. The horizontal lines represent term-specific means. This figure documents the consistent improvement in evaluations as class sizes have decreased, and the sharp shift in distribution in Fall 2007 with the new scale.

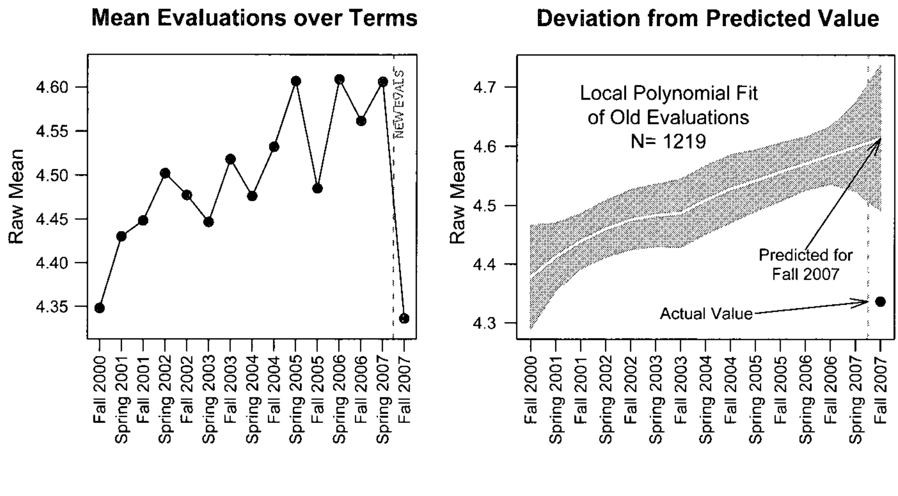


Figure 7: The left panel plots terms on the *x*-axis and conditional means on the *y*-axis. This panel shows an upward trend with a fall-spring effect, but a sharp drop in Fall 2007. The right panel plots predicted values from a local polynomial fit to the pre-Fall 2007 data. The actual Fall 2007 grand mean falls far below the predicted value extrapolating from the model.

*Confounding*

The primary threat to inference is that many other factors may be causing the drop in evaluations. The set of instructors teaching in Fall 2007, as well as the set of classes offered, may be unique. The student body surely is different. And class sizes are generally smaller. As a result, it remains difficult to cleanly attribute changes to the evaluation system.

Nonetheless, the drop in evaluations is sharp. The left panel of Figure 7 presents the mean evaluations over time demonstrating the strong upward trend (with a fall-spring effect) prior to Fall 2007. The mean evaluation was 4.61 in Spring 2006, but the mean plummets to 4.34 in Fall 2007. To account for time trends, we first fit a (locally weighted polynomial) regression using the pre-Fall 2007 data and calculate predicted values for all terms, including Fall 2007.[41] The predictions, with 95 percent confidence bands, are presented in the right panel of Figure 7. The observed value falls sharply below the predicted interval, further corroborating that something went awry in Fall 2007.

To account for confounding class sizes, Figure 8 subclassifies the data into enrollment quartiles.[42] Each of the four columns represents a quartile, and the

41. See Trevor J. Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction 171-72 (New York, 2001); and William S. Cleveland, Eric Grosse, and William M. Shyu, Local Regression Models, *in* Statistical Models in S 309 (John M. Chambers and Trevor J. Hastie eds., Pacific Grove, Cal., 1992).

42. See Daniel E. Ho, et al., Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference, 15 Pol. Analysis 199 (2007); and Sander Greenland, James M. Robins, and Judea Pearl, Confounding and Collapsibility in Causal

top panel plots the distribution of mean evaluations for pre-Fall 2007 classes and the bottom panel plots the distribution for Fall 2007 classes. At each level of enrollment, we observe a distributional shift, with the shift appearing slightly more pronounced for larger classes. These panels underscore that while the mean shift, identified by vertical lines, appears relatively small, the new evaluations induced fuller use of the 1-5 response categories.

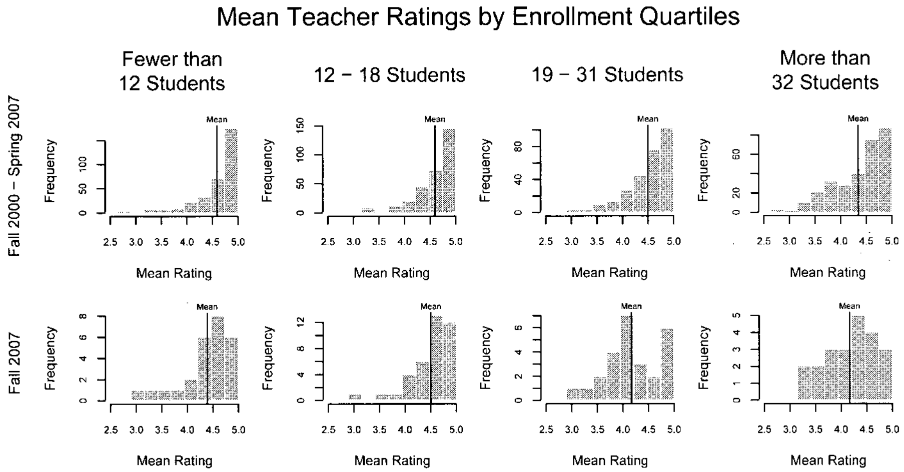


Figure 8: This figure plots histograms of outcomes in subclasses corresponding to quartiles of enrollment. The top panels present pre-Fall 2007 evaluations and the bottom panels present Fall 2007 evaluations. This figure shows a drop across each quartile, suggesting that the new evaluations affect all courses and that the effect is not confounded by class size. Vertical lines plot conditional means.

The most considerable bias of these estimates may be due to aggregation of different courses and instructors. To assess to what degree the effect may be driven by such compositional changes, Figure 9 shows the 67 exact course-instructor matches between Fall 2007 and the previous two semesters. We match course-instructor observations in the proximate terms to capture general time trends. For example, Professor Bankman's Fall 2006 Tax class is matched to the same class in Fall 2007, and Professor Daines's Fall 2006 Corporations sections are matched with the Fall 2007 sections. With this matched sample, differences are guaranteed not to be confounded by instructors and courses.[43] The effect remains considerable. The arrows in Figure 9 connect matches and are shaded increasingly darker for larger decreases or increases in mean scores. The width of the arrow is proportional to course enrollment to address sampling variability.

Inference, 14 Stat. Sci. 29 (1999). Pertaining to class size specifically, see also sources cited in *supra* note 40.

43. See Ho, et al., Matching as Nonparametric Preprocessing, *supra* note 42; Lee Epstein, et al., The Supreme Court During Crisis: How War Affects only Non-War Cases, 80 N.Y.U. L. Rev. 1 (2005); Daniel E. Ho, Why Affirmative Action Does Not Cause Black Students to Fail the Bar, 114 Yale L.J. 1997 (2005).

By and large, the mass of arrows points downwards. Several increases are driven by courses with small enrollment, which we would expect by chance alone. In short, holding constant course *and* instructor, we continue to see a decrease, as indicated by the mass of arrows in the figure.

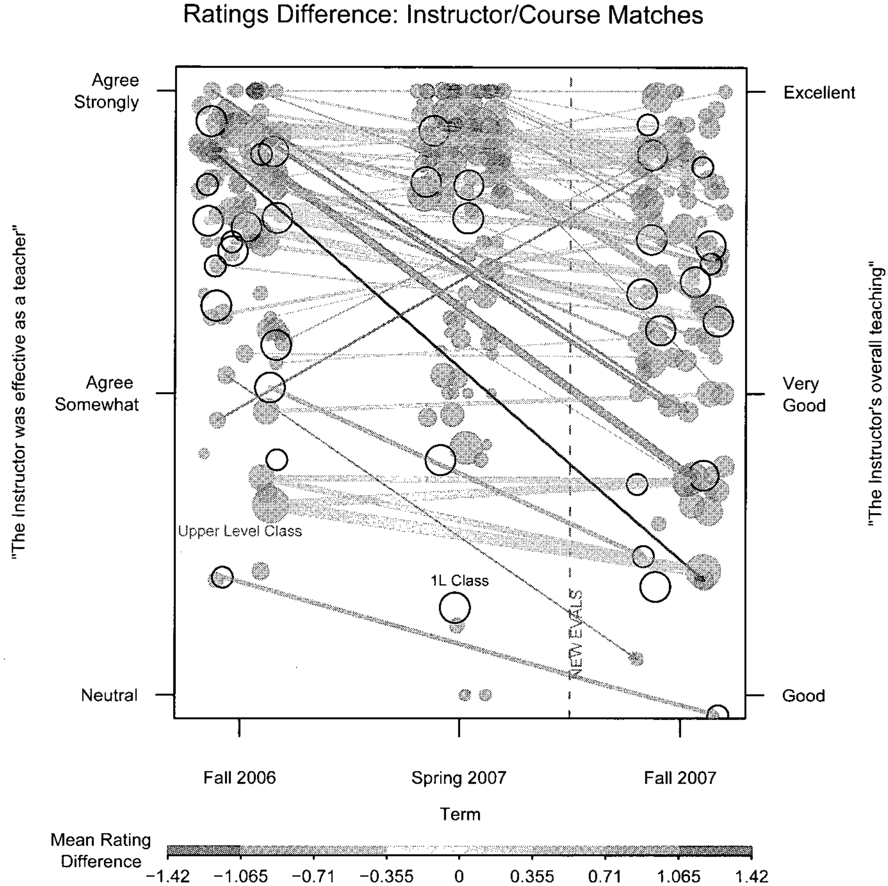


Figure 9: This figure plots the 67 exact matches of course-instructor units from Fall 2007 to Fall 2006 or Spring 2007. For example, Professor Bankman's Fall 2006 Tax class is matched with his Fall 2007 Tax class. Arrows represent the matches and are weighted by course enrollment. Darker arrows indicate a larger increase or decrease for a particular course-instructor match between semesters. Circles are weighted by course enrollment and horizontally jittered for visibility. This figure shows that overall evaluations decrease, holding constant course and instructor.

While Figure 9 shows that the effect does not appear driven by unique courses or instructors, dropping one-time instructors and courses may also not yield proper estimates of the impact of course evaluations. Dropping one-time instructors, for example, may bias the estimate downwards if long-time teachers are least susceptible to evaluation system changes.

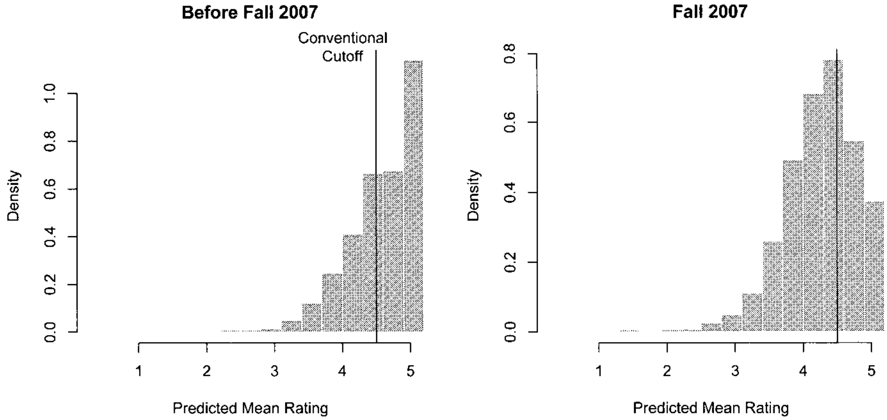


Figure 10: This figure visualizes the posterior predictive distribution for one selected course from the multilevel model detailed in Appendix B. The left panel models the predictions on the old evaluation system, and the right panel models the predictions on the new evaluation system. This figure shows that even with a small mean drop, the censoring at 5 and the variance shift cause large differences in whether a course falls above an informal 4.5 cutoff.

To account for these complex effects, we fit a simple multilevel model to the data, details for which are outlined in Appendix B.[44] The model has several key features. First, it accounts for the direct effects of enrollment, fall-spring differences, and a linear time trend (as validated by a nonlinear model in Figure 7). Second, the model accounts for instructor and course-specific ("random") effects. Third, it models both a mean and variance shift attributable to the evaluation change (one-tailed posterior *p*-values that there is no mean shift or that the variance is homogeneous ≈ 0). Lastly, it models the censoring of the outcomes, namely that many ratings bump up against 5.

Figure 10 presents posterior predictive draws for one actual course offered in Fall 2007, varying only the evaluation system (i.e., the mean and variance).[45] The left panel presents the predictions for the course on the old system, with a modal rating close to 5. The vertical line represents one conventional "cutoff" of 4.5, sometimes informally used as a performance check. The right panel presents the predictions for the *same* course on the new system. Both the variance and mean shift considerably: the modal rating is now around 4.5. While the mean shift isn't that large (roughly 0.24), the substantive effect, taking into account the variance shift and censoring, is considerable: for the same

44. See Andrew Gelman and Jennifer Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models (New York, 2007).

45. See Andrew Gelman et al., Bayesian Data Analysis (2d ed., Boca Raton, Fla., 2003); Andrew Gelman, Exploratory Data Analysis for Complex Models (with Discussion), 13 J. Computational & Graphical Stat. 755 (2004); and Andrew Gelman, Xiao-Li Meng, and Hal Stern, Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (with Discussion), 6 Stat. Sinica 733 (1996).

course, the old system yields a 35 percent probability that the instructor won't meet the cutoff, but that probability jumps to 59 percent with the new system. In short, the model strongly suggests that the new evaluations have caused a distributional shift in the ratings.

*Nonparametric Bounds*

Lastly, we investigate to what degree effects may be separated out into form versus substance. To do so, we focus on one upper-division course for which both paper and online evaluations were administered. This course appears to have been the only upper-division course in Fall 2007 for which paper evaluations were circulated on the last day of class, so as to accommodate students without laptops. Students in that class submitted forty-four evaluations online (though five were missing global ratings) and twenty-six evaluations on paper. Since only sixty-one students were enrolled, at least four students submitted duplicative ratings. At the outset, one concern for reporting results was whether to *disregard* the twenty-six paper evaluations, which would be correct only in the unlikely scenario that the paper evaluations were completely duplicative of the online evaluations.

Indeed, the mean online evaluations were 3.72 and the mean paper evaluations were 4.12, suggesting differences between the two forms. Figure 11 conducts a bounds analysis to relax unwarranted assumptions in combining these sources of information.[46] Rows 1 and 2 present point estimates on strong and unfounded assumptions (e.g., that paper evaluations add no information). Row 3 combines these two estimates using a weighted average, accurate only if individuals submitting multiple ratings are completely random. Rows 4 and 5 calculate bounds eliminating the top or bottom double-voters. Rows 6 to 9 use one source (paper or online) exclusively. Rows 6 and 7 present monotonicity bounds, under the assumption that missing evaluations are no worse or no better than the observed ratings. Rows 8 and 9 present fully nonparametric bounds under the worst-case scenario of all missing evaluations being 1 or 5.

46. See Charles F. Manski, Identification Problems in the Social Sciences 21 (Cambridge, Mass., 1996); and Joel L. Horowitz and Charles F. Manski, Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations, 84 J. Econometrics 38 (1998).

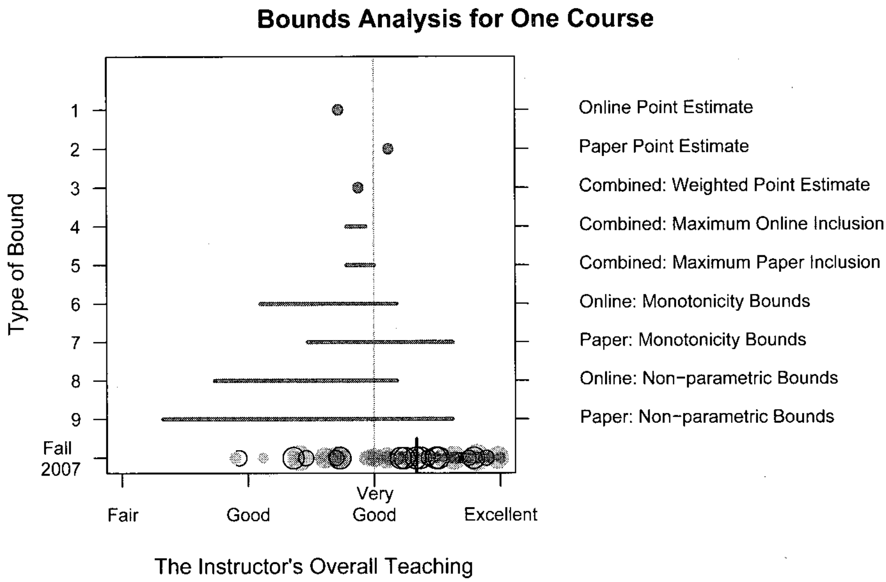


Figure 11: This figure presents a bounds analysis for one particular course that had forty-four online (with five missing ratings) and twenty-six paper evaluations for a class of six students. Row 1 presents the point estimate using only online evaluations. Row 2 presents the point estimate using only paper evaluations. Row 3 presents a weighted average. Rows 4 and 5 calculate bounds eliminating the top or bottom double-voters from the online or paper evaluations, respectively. Rows 6 to 9 do not combine sources of information (i.e., online and paper). Rows 6 and 7 plot bounds assuming that missing evaluations are no worse or better than observed evaluations. Rows 8 and 9 calculate strict nonparametric bounds, assuming that missing evaluations could be all 1s or all 5s. This figure illustrates the severe informational deficit as nonresponse increases.

This bounds analysis suggests two points. First, it provides suggestive evidence that the medium (paper versus online) may matter. That said, paper evaluations may have been submitted by different *types* of students. Even then, the difference suggests nonrandom nonresponse bias when online evaluations become the exclusive mechanism. Second, the bounds analysis shows the fragility of learning from student evaluations as the response rate drops. This is perhaps the most sobering challenge with the online system, as it appears to exacerbate (nonrandom) nonresponse.

## Implications

Our evidence strongly suggests that the simultaneous changes in wording, timing, and implementation of Fall 2007 affected the distribution of student evaluations. Meaningfully equating even seemingly similar ratings remains difficult. While we applaud the underlying rationale of attempting to gather more information about courses, the fact that the two scales are not anchored means that the law school may be *learning less* about courses and instructors even when *asking more* questions.

This study is the first, as far as we are aware, to document empirically these threats to comparability. We therefore conclude with several broader implications of this case study on the use, interpretation, and reform of teaching evaluations.

First, there is good news. Students reasonably respond to even subtle changes in questions asked of them. To meaningfully interpret evaluation results, then, one should interpret results in the context of the particular question asked. Summaries of evaluation responses such as those presented in Figure 2 are far more desirable than naive numerical means that provide superficially similar scales. The visualization techniques we illustrate have also uncovered broad dynamic trends, such as fall-spring effects and long-term upward trends in evaluations, which can empower the interpretation of student evaluations, as well as shed light on broader institutional trends.

Second, any institution considering reform of its teaching evaluation system should do so with scale equating in mind. Standard approaches are available to calibrate scales (such as by random assignment of questions and/or forms). A form of stratified randomization, where half of the students enrolled in a course are given one form and half the other, would be easy to implement and would facilitate reliable, robust scale equating.

Third, our analysis empirically confirms that online evaluations are more prone to nonresponse than traditional in-class evaluations. Nevertheless, online systems present tangible benefits in cost reduction and ease of analysis. To reduce nonresponse but exploit the many advantages of online evaluations, we suggest that paper evaluations should still be made available to students, especially as certain instructors have discouraged (if not banned) laptops from the classroom. To prevent "double-counting," a system similar to Axess's student ID verification could be implemented with paper evaluations, thereby maximizing integrity and eliminating effects caused by format-specific perceptions of anonymity.

Lastly, to maximize comparability and minimize timing effects, evaluation reform should strive to hold constant the timeframe of submission. Constructing online surveys has become cheaper and easier, and with many online firms offering flexible customized forms and delivery methods, shortening the submission window should be fairly easy. With timeframe changes as drastic as those observed in this case study, evaluations post-implementation are transformed into something wholly different from before.

Like it or not, teaching evaluations retain a critical role in legal education. Our article places the understanding, interpretation, and improvement of evaluations on firm empirical ground.

## Appendix

### *Data Collection*

Our analysis is based on course-level aggregated data obtained from Stanford's Office of Student Affairs. The data came in two parts: course-level statistics for (1) all classes from Fall 2000 to Spring 2007 and (2) Fall 2007 first-year classes, which were still administered in paper on the last day of class. We augmented this data with data from upper-level Fall 2007 course evaluations, the first to be submitted online, through Stanford's Axess website. These data were cleaned (removing commas, notes, and text formatting) and compiled as a uniformly formatted dataset spanning the entire fifteen semester observation period.

We accounted for a number of data inconsistencies. First, the raw data exhibited considerable variation in course titles across semesters. For example, the "Law and Economics Seminar" was listed in different terms as "Law & Econ Sem," "Law & Economic Seminar," and "Law and Econ Seminar." Similarly, instructor names varied considerably. For example, Professor A. Mitchell Polinksy was referred to as: "Polinsky, A Mitchell," "Polinsky, Mitch," "Polinsky, A.," and "Polinsky, Mitchell." Each of inconsistencies was manually cleaned to generate uniform course and instructor IDs.

Second, matching evaluations with instructors proved challenging due to inconsistencies in recording evaluations of co-taught courses. When courses included separate instructor ratings for each teacher (e.g., separate Supreme Court Clinic entries for Professor Pamela Karlan and Thomas Goldstein), entries were treated as separate units. Co-taught courses that contained only one mean rating were assigned a unique ID (for the pair or trio of instructors), as it remains unclear to whom individually the ratings correspond.

Third, the enrollment and response data for thirty-nine courses was clearly mistaken, yielding a response rate higher than 100 percent (a logical impossibility). For example, the response rate for Law and Science of California Coastal Policy in Spring 2007 was 425 percent. We assume that enrollment and response numbers were simply switched.

Fourth, courses with either a 0 percent response rate or missing instructor evaluations were discarded. These situations arose when instructors forgot to hand out evaluations or when instructors elected to use other methods of evaluation for specific classes.

Table 3 summarizes the number of units affected by these recodings.

| Recoding Issue | Number | Proportion of Data Affected |
|---|---|---|
| Inconsistent course names | 200 | 15.03% |
| Inconsistent instructor names | 104 | 7.81% |
| Co-taught courses | 25 | 1.88% |
| Enrollment lower than responses | 39 | 2.93% |
| Missing data | 32 | 2.41% |

Table 3: Cleaning and recoding of raw evaluation data.

*Parametric Adjustment*

To simultaneously adjust for many of the confounding factors, we use a Bayesian multilevel model.[47] The model can be written as:

$$Y_{ij}^{*} \sim N(T_{ij}\tau + X_{ij}\beta + \gamma_i + \delta_j, \sigma(T_{ij}))$$

where N() is the normal distribution, $i$ indexes classes, $j$ indexes teachers, $T_{ij}$ equals one if course $i$ is taught by instructor $j$ in Fall 2007, $X_{ij}$ includes enrollment, the year, and an indicator for fall, and $Y_{ij}^*$ is a latent variable to account for censoring at 5. The key identifying assumption, similar to type of regression-discontinuity design, is that the mean trend is locally linear. The observation mechanism is:

$$Y_{ij} = \begin{cases} Y_{ij}^{*} & \text{if } Y_{ij}^{*} \leq 5 \\ 5 & \text{otherwise} \end{cases}$$

The variance $\sigma$ $(T_{ij})$ is allowed to differ for the pretreatment period and Fall 2007. The model accounts for instructor and class random effects, assumed to be drawn from common hyperdistributions:

$$\gamma_i \sim N(\mu_\gamma, \sigma_\gamma)$$
$$\delta_j \sim N(\mu_\delta, \sigma_\delta)$$

We assume diffuse priors for remaining parameters, and use Gibbs sampling to draw a sample of 1,000 draws from the joint posterior. We use R and WinBUGS to fit the model.[48] Standard diagnostics suggest convergence. The 95 percent posterior interval for the mean parameter $\tau$ is (-0.44, -0.23), with a posterior probability of roughly 100 percent that the mean shift is negative.

47. Gelman and Hill, Data Analysis, *supra* note 44.

48. R Core Development Team, R: A Language and Environment for Statistical Computing (Vienna, 2007) (available at <http://www.R-project.org>(last visited Jan. 9, 2009)); Sibylle Sturtz, Uwe Ligges, and Andrew Gelman, R2WinBUGS: A Package for Running WinBUGS from R, 12 J. Stat. Software (3) 1 (2005); and David J. Lunn et al., WinBUGS–A Bayesian Modelling Framework: Concepts, Structure and Extensibility, 10 Stat. & Computing 325 (2000).