

Changing incentives in a multitask environment: evidence from a top-tier business school

James A. Brickley^{*}, Jerold L. Zimmerman

*William E. Simon Graduate School of Business Administration, University of Rochester, Rochester,
New York 14627, USA*

Accepted 22 June 2001

Abstract

This study focuses on changes in incentives at the William E. Simon Graduate School of Business Administration in the early 1990s to redirect effort from academic research to classroom teaching. We find a substantial and almost immediate jump in teaching ratings following the changes in incentives. Longer-run learning and turnover effects are present. Evidence also suggests that research output fell. This case illustrates the power of organizational incentives to redirect effort in a multitask environment, even in the presence of apparent human–capital constraints. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Incentives; Business schools; Research

1. Introduction

Agency theory suggests that the principal is interested in both the amount of effort exerted by the agent, as well as the agent's allocation of effort across tasks.¹ As environments change, firms are expected to adjust incentive contracts on both dimensions. For example, the 1990s witnessed significant developments in information technology, which lowered the costs of measuring performance. These

^{*} Corresponding author. Tel.: +1-716-275-3433; fax: +1-716-442-6323.

E-mail address: brickley@simon.rochester.edu (J.A. Brickley).

¹ For example, see Holmstrom and Milgrom (1991) for a general model and Feltham and Xie (1994) and Hemmer (1995) for specific modeling applications.

changes potentially help to explain why many firms increased their use of incentive compensation over this period. Similarly, changes in competition and technology motivated numerous firms to increase their focus on quality, for example, through the adoption of TQM programs (Wruck and Jensen, 1994; Brickley et al., 2001).

Empirical work on the effects of changes in incentive contracts is limited. Papers such as Banker et al. (1996), Brown (1992), and Lazear (1996) provide evidence supporting the hypothesis that increases in incentive pay have a significant and rapid effect on effort and output. To our knowledge there is no empirical evidence on how changes in incentives motivate workers to alter their allocation of effort across tasks.

A priori, one might expect that the existing stock of human capital limits organizations' abilities to alter workers' outputs across tasks in the short run. If a firm has selected and trained its work force based on one primary dimension, such as the ability to produce high output, its work force likely lacks the skills, ability, and/or interest to quickly shift to some other dimension, such as quality. Such a change requires time to train and replace employees with ones with the requisite skills. Thus, human capital is a potentially important source of "organizational inertia"—forces inside an organization which make it resistant to change.

In this paper, we study changes in incentives at the William E. Simon Graduate School of Business Administration (University of Rochester) during the 1990s. The School's policies relating to the selection and retention of faculty, as well as training and mentoring had long emphasized research. Indeed, it is fair to say that teaching received little weight in the personnel decisions for tenure-track faculty in the 1970s and 1980s. During the early 1990s, there was a substantial environmental shift that increased the importance of teaching relative to academic research at top business schools. The Simon School, like other business schools, made changes in its performance evaluation and reward systems to increase the emphasis on teaching.

Our evidence indicates that despite apparent human capital constraints, the faculty adjusted very rapidly to the new incentive structure. Average teaching ratings increased from about 3.8 to over 4.0 (scale of 5) surrounding the year when the incentive structure was changed. This increase was observed throughout the School, across types of faculty (tenure track and non-tenure track) and programs (e.g., regular and executive MBA programs). Teaching ratings continued to rise after the changes in incentives, suggesting additional learning and turnover effects. Intense peer review of classes, however, had no direct effect on teaching ratings for either the evaluated classes or subsequent courses taught by the instructor. Consistent with the multitask agency model's prediction of faculty substituting effort from research to teaching, we present evidence that suggests research output fell after more incentive was placed on teaching.

It is important to note that the administration and faculty at the Simon School *chose* to increase the strategic emphasis on teaching. Presumably, they based this

decision, in part, on the ability of the faculty to adjust to the changes in incentives. Therefore, the speed with which the Simon faculty was able to adjust is potentially higher than a school where the administration and faculty decided it did not make sense to increase the focus on teaching. Nevertheless, the very rapid adjustment at the Simon School highlights the power of incentives in affecting the allocation of effort across tasks.

Our work is similar to past studies on organizational incentives (e.g., Banker et al., 1996; Lazear, 1996) in that we focus on a large data set from one organization. This approach is common in the literature due to the lack of publicly available data on compensation and incentive policies. One advantage we have in this study is that we are long-time employees of the organization. Our knowledge of the organization's history allows us to control for confounding factors that the typical "outside" researcher might not even know existed. Also, while we study a single organization, there is significant heterogeneity in the data (places of instruction, type of class, program, etc.) which increases the power of our tests.

The next section describes the underlying environmental factors that caused the Simon School to re-evaluate the incentives for teaching and research and how the school changed the way teaching was measured. The paper's five hypotheses are discussed in Section 3. Section 4 describes the data and Section 5 presents the results of the tests. The last section reviews the paper's principal findings.

2. Background

2.1. Environmental factors and strategic shifts

By the late 1980s and early 1990s, domestic demand for MBAs was falling, due at least in part to the decline in the birth rate. Business schools began competing more intensively for students by spending more on public relations and admissions and by giving more scholarships. Contributing to the competition and helping to shape it, *Business Week*, began publishing in 1988 a biannual list of the top 20 business schools. Graduating students and recruiters rated the schools. The survey gave no explicit recognition of faculty research contributions. Students were asked to judge the opportunities provided by their schools "— either in class or in extracurricular activities—to nurture and improve your skills in leading others."² The Simon School made the top 20 list in 1988 and 1990, but not in 1992.

Competition among business schools led faculties to reassess their strategic plans. For example, MIT Sloan School issued a report in May of 1991 that, among

² J. Byrne and D. Leonhardt, "The Best B-Schools," *Business Week* (October 21, 1996).

other things, called for enhanced incentives to improve teaching.³ A similar process in Rochester recommended to “increase teaching incentives, and make the change clearly visible to applicants, students, administrators and faculty.”⁴ Improving teaching was viewed as a way to immediately increase the *Business Week* ratings as well as enhance the School’s ability to attract high quality students.

2.2. Evaluating teaching

Complicating the Simon School’s ability to alter incentives was the fact that a faculty member’s teaching effectiveness (output) is largely unobservable. It has both quantity (number of students and classes taught) and quality dimensions. Quality includes unobservables such as the breadth and depth of topics covered and the amount of understanding retained by the students. Teaching also includes increasing the capital stock of pedagogy such as cases, textbooks, lecture notes, slides, and problem material that other instructors can use. “Good” teachers help their colleagues become better teachers. They also make the learning experience exciting for their students.

2.2.1. Teaching ratings

Like most organizations, business schools measure teaching performance using easy to observe variables, such as student surveys. At the end of each course, Simon School students complete an evaluation form. They answer two questions using a 1 to 5 scale: “What is your evaluation of the instructor?” and “What is your evaluation of the course?” The students also describe the strengths and weaknesses of the course. The Dean’s office reviews the completed forms. The mean scores on the two questions are distributed to the entire faculty.⁵ The original forms with the written comments are returned to the instructor. Prior to 1992, the Simon School used student teaching ratings as its primary if not sole measure of teaching effectiveness.

During the early 1990s the School increased its emphasis on teaching ratings in a number of ways. First, the School began awarding a small research stipend to the five faculty receiving the highest student ratings each quarter and placing their names on a prominently displayed plaque. Second, the Associate Dean began

³ “The Report of the Task Force on Improvement,” MIT Sloan School (May 7, 1991).

⁴ “MBA Program Status Report,” University of Rochester William E. Simon Graduate School of Business Administration (June 14, 1991).

⁵ Note that since student evaluations are an ordinal measure, the mean rating is a questionable construct. The numbers 1, 2, 3, ... are only used as orderings. Below, we supplement traditional parametric tests with medians and other statistical methods that do not violate the ordinal nature of the evaluations.

indicating that he reads the student evaluations by penning a short note to instructors. This note is included with the tabulated student ratings.

2.2.2. CTE reviews

Student ratings provide a quantitative measure of teaching effectiveness. Previous research, however, indicates that student teacher ratings and learning are only weakly related. Gramlich and Greenlee (1993) study over 15,000 economics students who are graded in common exams and are taught from teachers who each receive a student evaluation. Using the common exam as a measure of student learning they find that teacher ratings are slightly related to grades.

Another concern is that some instructors gave student ratings by reducing course work loads and cutting analytic content.⁶ Some instructors hand out cookies, bagels, and wine and cheese the last day of class when student ratings are administered. Student course ratings are least useful as a measure of course design and development efforts because students lack a reference base of similar courses at other schools (Calderon et al., 1996; Martin, 1998).

These concerns about teaching ratings motivated the Simon School faculty to pass the following proposal in Winter 1992:

To establish a faculty committee to evaluate teaching content and quality on an on-going basis. The intent of the proposal is to put the evaluation of teaching on the same footing as the evaluation of research. The committee will have the responsibility to evaluate both the content and presentation of each faculty member on a regular basis to be determined by the committee. . . . The output of this process should be reports designed to provide constructive feedback to faculty and evaluations to be considered in promotion, tenure, and compensation decisions.⁷

To implement this proposal, the Dean's office formed The Committee on Teaching Excellence (CTE) in the spring of 1992. The nine faculty members on the CTE developed a set of procedures by evaluating six courses currently taught by members of the CTE during the 1993 academic year (fall quarter 1992 through summer quarter 1993).

By the end of the 1993 academic year, the CTE established a process, that except for minor changes, remained in effect through 1997. This process includes

⁶ Aigner and Thum (1986) report that hours per week required outside of class had a statistically significant negative impact on teacher ratings by undergraduate economics students.

⁷ "Faculty Meeting Minutes," University of Rochester William E. Simon Graduate School of Business Administration (February 26, 1992).

benchmarking the class with other top business schools; using a two-person evaluation team to observe lectures, review material, and conduct student focus groups; video taping several classes; full committee discussion of the course; and a final written report which goes to the instructor and the Dean's office and which is included in the faculty member's personnel file. These CTE evaluations resemble referee reports in that they discuss strengths, weaknesses, and suggested changes. No numerical or quantitative rating is assigned. This process is very time intensive. The former CTE chair estimates the opportunity costs to evaluate one course at US\$15,000.

In addition to evaluating nine individual courses each year, the CTE held several seminars to discuss teaching. These forums allowed faculty to share their experience on various topics including: teaching cases, using computer-based presentation packages, and managing class discussion ("cold" calling). These seminars in the 1995 academic year were the first faculty seminars devoted to teaching.

2.3. Evaluating research

Faculty research is reviewed annually by the Dean's office as part of the salary adjustment process and periodically for contract renewals or promotion decisions. Annually, faculty members submit an activities report that categorizes their research by: "List publications which have appeared during the year," "List working papers from previous years and indicate status of each publication (e.g., articles accepted, under revision, or submitted for publication)," and "List working papers which have appeared for the first time during the last year." The Dean's office does not read faculty research papers as part of the year-end evaluation process, but rather relies on the School's small size and word of mouth to provide them with information regarding the importance of various faculty research. In addition, the Deans often attend faculty research seminars during the year.

For contract renewals and promotion and tenure decisions, faculty committees read all the candidate's research and thoroughly discuss it in the presence of the Associate Dean before making a recommendation to the Dean's office. Promotion decisions require outside letters regarding the candidate's research record. In making personnel decisions, the weights placed on research, teaching, and service vary with the level of the promotion decision. Prior to the increased teaching incentives, assistant professor contract renewals were based almost exclusively on research potential. Teaching had to be "tolerable." Virtually no weight was placed on school service. For tenure decisions, the research must be recognized as having made a significant contribution to the field, while teaching and service must be acceptable. Following the changed incentives, research continues to be most important in the evaluation process. However, more emphasis is placed on teaching at all levels of review. Good teaching can also generate opportunities to

continue working at the School for a period of time when there has been a negative promotion or tenure decision.

2.4. Rewarding teaching and research

While most Simon faculty agree that teaching incentives increased at the School since the early 1990s, there was no *explicit* change in the School's performance reward policy (other than the new nominal teaching awards). For instance, following the formation of the CTE in 1992, the Dean's office did not issue a statement linking salary adjustments directly to CTE reports.⁸

Annual salary adjustments are the sole province of the Dean's office. In revising salaries, the Dean and Associate Dean review each faculty's research, service, and teaching evaluations as measured by the student evaluations. The Deans use the average student teacher rating for all courses taught by the instructor in the last 12 months. In the salary letters written to the faculty, it is not uncommon for the Dean to refer to exceptionally good or bad student teaching ratings. This procedure for adjusting salaries and for informing faculty of their salaries has been in effect since 1985.

When deliberating on promotions, the Promotion and Tenure Committee receives a report containing a time series of the student ratings on all courses taught by the candidate. The CTE report on the candidate is also included in the packet of material reviewed by this committee.

In summary, the School does not base salaries or promotions on a mechanical formula that includes student teacher ratings or CTE reports or on the number of papers published. Faculty salaries and promotions are based on subjective performance evaluation where the weights for teaching and non-teaching activities are unstated. Moreover, the weight placed on the CTE report relative to the student ratings by the deans in making salary and promotion decisions is not known by the faculty. However, since student ratings are calculated for each class taught whereas CTE reports are prepared at most every 5 years, student ratings appear to be the more important of the two performance measures.

Fig. 1 presents a chronology of the events preceding and following the formation of the CTE. Based on this time line, we focus on 1992 as the academic year (Fall 1991–Summer 1992) in which the Simon School began to increase

⁸ One potential criticism of this study is that in contrast to Banker et al. (1996) and Lazear (1996) we study an organization where there was no *explicit* change in compensation policy (while the changes in incentives were arguably strong, they were *implicit*). As discussed below, the Simon School did make an explicit adjustment in the compensation for doctoral students during our study period, linking pay to teaching performance. This fact allows us to check our basic results using a sample where incentives were explicitly changed.

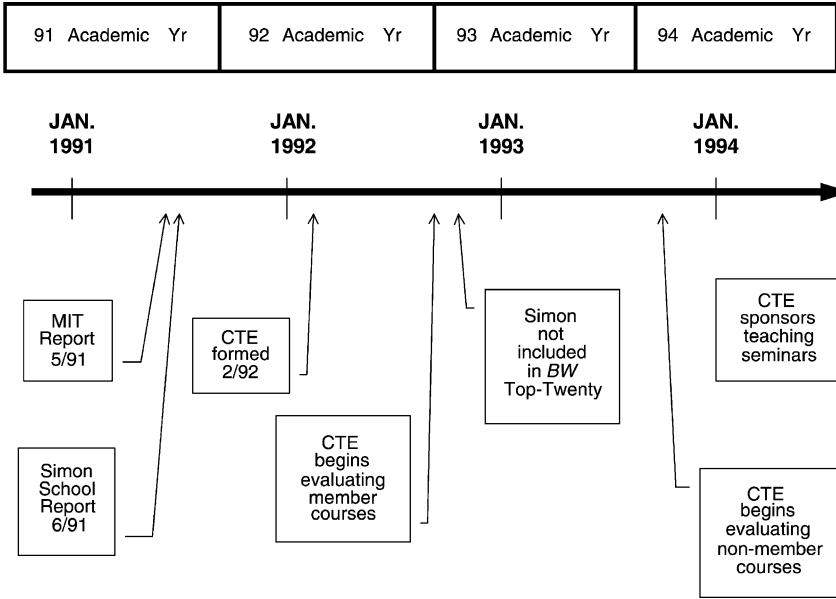


Fig. 1. Chronology of significant events.

incentives for teaching.⁹ The MIT and Simon School reports (which called for increased teaching incentives) were produced at the end of the 1991 academic year (May 1991 and June 1991). The proposal to form the CTE was formulated at the beginning of the 1992 academic year and was adopted by the faculty in February of 1992.

3. Hypotheses

Subsections 3.1 and 3.2 describe hypotheses regarding how teaching evaluations changed with the introduction of the CTE. Subsection 3.3 discusses our predictions regarding the change in research output.

3.1. Immediate incentive effects on teaching (Hypotheses 1 and 2)

Faculty allocate effort across a variety of tasks: research, teaching, service, and consulting. Multitask agency theory predicts that the change in incentives at the

⁹ Choosing 1992 as the “breakpoint” is not critical for our analysis. For instance, similar results hold if we treat 1992 as a transition year and 1993 as the breakpoint.

Simon School beginning with the 1992 academic year would have motivated faculty to increase the effort devoted to teaching and reduce effort devoted to other activities such as research (Holmstrom and Milgrom, 1991). The first prediction is tested using regressions of the following general form:¹⁰

$$S_i = a_1 + a_2 \text{POST91}_i + a_3 \text{CTE}_i + a_4 X_{1i} + a_5 X_{2i} + \dots + u_i \quad (1)$$

where S_i = mean student rating of course i (ranges between 1 and 5); $\text{POST91}_i = 1$ if the i th course is taught in academic year 1992 or later, 0 otherwise; $\text{CTE}_i = 1$ if course i was evaluated by the CTE in the given quarter, 0 otherwise; X_{ji} = control variables, $j = 1, 2, \dots, n$.

We are interested in testing two hypotheses: (i) introducing the CTE sent a credible signal to the faculty that teaching was now more important and would be supported by the School and in response, faculty increased the amount of effort devoted to teaching ($a_2 > 0$), and (ii) peer review of specific courses caused student ratings of these courses to increase ($a_3 > 0$). The first hypothesis predicts that all student course ratings increase following the establishment of the CTE. The second hypothesis predicts that faculty effort on specific courses increase when they are evaluated by the CTE. The two hypotheses are not mutually exclusive and each is discussed in greater detail below.

3.1.1. CTE signals increased importance of teaching ($a_2 > 0$)

Prior to forming the teaching committee, numerous statements by the deans and faculty implored the importance of teaching. However, such statements lacked credibility. Given the subjective nature of performance evaluation and reward, faculty were uncertain about the return from various tasks including teaching. Most retention, promotion, hiring, and compensation decisions suggested that research was by far the most important task at the School. It was not until the School, and in particular the senior faculty, endorsed the formation of the CTE that faculty came to believe that the return to teaching had increased. By agreeing to devote substantial real resources to evaluating teaching, the Dean's office and the senior faculty sent a credible signal (inference) that the importance of good teaching had increased.

3.1.2. Peer review increases student ratings ($a_3 > 0$)

When the CTE evaluates a particular course, the evaluatee is predicted to shift additional effort from other courses and activities such as research, service, and consulting to the course being evaluated and that this additional effort shows up in

¹⁰ The CTE variable in this regression is potentially endogenous. This issue is addressed in the empirical analysis.

teacher ratings. If the CTE values various course attributes such as work load and course development that students do not value, then additional faculty effort will not increase student evaluations. Testing this hypothesis using Eq. (1) also assumes that faculty expect additional rewards in the form of pay or promotion resulting from favorable CTE reviews. Both of these issues are discussed below.

3.2. Long-term learning and turnover effects on teaching (Hypotheses 3 and 4)

Theory also suggests that teaching ratings would continue to rise after 1992 because of the increased incentives of faculty to acquire additional human capital to improve their teaching. Also changes in the hiring, promotion, and tenure standards of the School would be expected to produce a faculty with higher teaching skills. To test this third hypothesis we examine trends in teaching ratings after 1992.

Student ratings would also increase over time if there are important learning effects that accompany the CTE evaluation process. For example, faculty might become better teachers through feedback and interaction with the CTE or through sharing ideas on how to present material, use audio-visuals, conduct case discussions, and so on. We test this fourth hypothesis by examining whether teaching ratings are higher in classes taught by instructors *subsequent* to their CTE evaluation.

3.3. Substitution away from research (Hypothesis 5)

If the change in incentives at the Simon School motivated faculty to increase the effort devoted to teaching, then the effort devoted to other activities such as research should fall. Hypothesis 5 predicts that less research is produced following the increased teaching incentives than before the change. Three factors, however, might mitigate this effect. First, faculty might substitute effort devoted to service or consulting instead of research, especially if they enjoy doing research. Second, if faculty believe that external markets for their services continue to be driven primarily by research and not by teaching, then their incentive to substitute research effort for teaching is reduced. Finally, the non-pecuniary income from presenting papers at conferences and at other schools increases the likelihood of rejecting Hypothesis 5.

4. Data

4.1. Teaching evaluations

The teaching evaluations data set consists of 3129 courses spanning the time period Fall quarter 1982 through Spring quarter 1997. For each class, two

questions are asked, “What is your evaluation of the instructor?” and “What is your evaluation of the course?” Both use a five-point scale: 1 = poor, 2, 3, 4, 5 = excellent. A final sample of 2227 classes result after deleting lab sections, non-credit courses, and doctoral and undergraduate courses. We also collected data on faculty characteristics such as full vs. part-time, tenured vs. untenured, and tenure track vs. non-tenure track.

Table 1 provides descriptive statistics on the course ratings as well as some of the control variables used in subsequent tests. Over the entire sample period the average instructor (INSTRUCTOR) is rated 3.98 with a standard deviation of 0.61. The course with a mean of 3.77 is rated on average lower than the instructor. The CTE evaluated 2% of the courses in the data set. The average number of evaluations per course (SIZE) is 33.33 students. The Simon School offers both a full-time and part-time MBA. Part-time MBA courses are taught in the evening. Half the courses are NIGHT sections. The School also offers three executive MBA degree programs, one in Rochester and two in Europe. An Australian program, started in 1991, stopped in 1993 due to lack of demand. EDP denotes executive MBA classes taught in Rochester and IEDP denotes executive classes taught abroad. Domestic and foreign executive MBA classes each constitute about 10% of the total courses.

Three dummy variables are used to characterize instructors. About 83% of the classes are taught by full-time faculty (FULLTIME), 66% are taught by faculty on

Table 1
Descriptive statistics (maximum number of observations = 2227)

Variable	Mean	Std. Dev.	Minimum	Maximum
INSTRUCTOR	3.98	0.61	1.45	5.00
COURSE	3.77	0.55	1.6	5
CTE	0.02	0.15	0	1
SIZE	33.33	18.60	2	131
NIGHT	0.50	0.50	0	1
EDP	0.10	0.30	0	1
IEDP	0.09	0.28	0	1
FULLTIME	0.83	0.38	0	1
TENTRACK	0.66	0.48	0	1
TENDUM	0.28	0.45	0	1

INSTRUCTOR: Mean student rating of instructor.

COURSE: Mean student rating of course.

CTE: 1 if course is evaluated by the CTE, 0 otherwise.

SIZE: Number of student evaluations.

NIGHT: 1 if course is taught in the evening, 0 otherwise.

EDP: 1 if course is U.S. executive MBA, 0 otherwise.

IEDP: 1 if course is an international executive MBA, 0 otherwise.

FULLTIME: 1 if instructor is full time, 0 otherwise.

TENTRACK: 1 if instructor is on the tenure track, 0 otherwise.

TENDUM: 1 if instructor is tenured, 0 otherwise.

the tenure track at Rochester (TENTRACK), and 28% of the classes are taught by tenured faculty at Rochester (TENDUM).

Not reported in Table 1, the number of classes taught per year more than doubled from about 80 in 1983 to over 200 in 1997. The increase is monotonic over the sample period and in part results from starting the foreign executive programs in 1988, 1991, and 1995. Also, the size of the entering class of full-time MBA students increased roughly 50% over the sample period. The new sections were staffed by a combination of additional full-time and adjunct faculty and extra compensation for existing faculty teaching overloads.

The heavy solid line in Fig. 2 plots the mean instructor rating over time. Consistent with our first hypothesis, there appears to be an upward shift in instructor ratings beginning in academic year 1992. The dotted horizontal lines at 3.87 and 4.07 represent the mean ratings for 1982–1991 and 1992–1997, respectively. We report the statistical significance of this time-series shift below. Our purpose here is to dispel some alternative explanations for the time-series shift as well as to provide a general sense of the data.

The increase in average instructor rating could be due to changing composition of either the faculty or students over this period. For example, the number of courses taught increased as did the percentage of international students. Fig. 2 provides two additional time series. The average instructor rating of tenure-track faculty (the light solid line) shows a similar trend as the entire sample. Hence, the

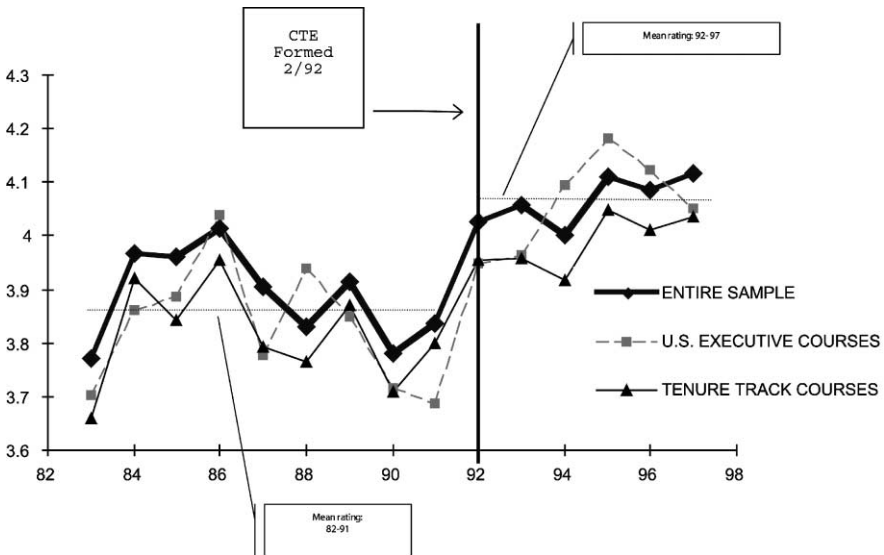


Fig. 2. Time-series of instructor ratings: entire sample, U.S. executive courses only, and tenure track faculty only.

Table 2A

Comparison of teaching ratings for the periods 1990–1991 and 1993–1994. Mean teaching ratings for each subperiod by type of faculty

Type of faculty	# of classes	% of total	Mean 1990–1991	Mean 1993–1994	<i>t</i> -Stat (difference) ^a
All Faculty	705	100	3.81	4.03	4.65
All Part-time (adjuncts)	106	15	3.78	4.19	3.73
All Full Time	599	85	3.81	4.00	3.48
Non-Tenure Track	132	19	4.04	4.18	1.23
Tenure-Track	467	66	3.76	3.94	3.12
Tenured	196	28	3.91	3.99	0.81
Untenured/ Tenure-Track	271	38	3.66	3.90	3.19

^aNull hypothesis: The mean teaching ratings for 1990–1991 and 1993–1994 are equal.

time-series shift is not due to hiring popular adjunct faculty. Likewise, the average teacher ratings in the domestic executive development program (the dashed line) also evidences an increase in ratings. This subset of courses holds relatively constant the student population, as well as the courses taught.

Table 2(A) and (B) presents additional information highlighting the substantial increase in teaching ratings around 1992. Table 2 (Panel A) compares the average teaching rating in 1990 and 1991 with the average rating for 1993 and 1994 for each of the major faculty classifications. For the full sample, the mean rating in the 1990–1991 period is 3.81 vs. 4.03 for the 1993–1994 period. An increase in mean teaching ratings is indicated for each group (adjuncts, all full-time faculty, tenure-track faculty, tenured faculty, and untenured faculty who are on the tenure track). Most of the differences are significant at the 0.01 level. Interestingly, the

Table 2B

Comparison of teaching ratings for the periods 1990–1991 and 1993–1994. Distribution of individual student ratings by subperiod^a

Rating	Number 1991–1992	% 1991–1992	Number 1993–1994	% 1993–1994
1	527	4.2	353	2.7
2	1166	9.3	829	6.3
3	2632	20.9	2139	16.4
4	4423	35.1	4803	36.8
5	3837	30.5	4930	37.8
Total	12,585	100.0	13,054	100.00

χ^2 test for independence of rows/columns: 285.7 (4 *df*); *p*-value = 0.00.

^aIn this panel, the individual student rating is treated as the unit of observation (not the mean of ratings for the class).

group that had among the highest increases in mean teaching ratings is the untenured, tenure-track faculty. This group which consists primarily of assistant professors arguably experienced the biggest change in incentives within the School. Prior to 1992, assistant professors were evaluated almost entirely on research output. Tenured faculty were judged on a broader set of criteria, including service to the School and teaching. Faculty not on the tenure track have always been evaluated primarily on teaching. During 1990–1991, the mean teaching rating was significantly higher for tenured faculty (3.91) compared to untenured faculty on the tenure track (3.66) ($t = 2.96$). There was no significant difference in these groups during the 1993–1994 period ($t = 1.12$). Table 2 Panel B reports a non-parametric test that the distribution of student evaluations shifted surrounding the creation of the CTE. There were fewer low student evaluations (“1,” “2,” and “3”) and more high evaluations (“4” and “5”) after 1992.

4.2. Research output

We use two measures of the “quantity” of research: the number of self-reported working papers by each faculty member in their annual activity reports and the number of new papers added to the School’s working paper lists. The first data set likely overstates research output. Faculty tend to list in their activity reports very

Table 3
Frequency distribution of number of new papers reported in activity reports and listed in working papers series per faculty per year, 1985–1995

Number of working papers	Reported papers		Listed papers	
	No. of observations	% of observations	No. of observations	% of observations
0	92	23.7	186	48.1
1	112	28.8	76	19.6
2	79	20.3	57	14.7
3	44	11.3	25	6.5
4	24	6.2	17	4.4
5	14	3.6	9	2.3
6	13	3.3	6	1.6
7	2	0.5	6	1.6
8	4	1.0	1	0.3
9	1	0.3	1	0.3
10	2	0.5	0	0
11	1	0.3	2	0.5
12	1	0.3	1	0.3
	389		387	

Reported Papers: number of new working papers self-reported in the annual faculty activity report.

Listed Papers: number of new papers listed in the official working paper series during the year.

early first drafts of papers, sometimes before they are ready for circulation. The second measure, papers listed in the working paper series, tends to be an understatement of research output. Once in the working paper series, anyone can request a copy. Faculty tend to place papers in the working paper series only after they are quite polished and ready for submission to a journal. Many papers that are listed on faculty activity reports as new working papers never are listed in working paper series. Both data sets are noisy measures of annual research output because much of the research could have been conducted several years earlier. Also, they do not control for changes in the quality, importance, or potential impact of the research. Given these caveats, we have data for 56 tenure-track faculty for the calendar years 1985–1995. The Dean’s office collects the data by calendar years and not by academic years.

Table 3 presents the distributions of the number of new working papers reported in activity reports and added to the working paper series. Both the median and modal response of reported papers was one new working paper during the year, with 28.8 percent of the observations. The most number of reported papers was 12. The median number of papers listed in the working paper series was also 1 for the year, but the mode was zero.

Fig. 3 plots the two time series of the average number of new reported and listed papers per calendar year (January–December). The reported data (the solid line) does not reveal a significant discrete shift in research productivity in the

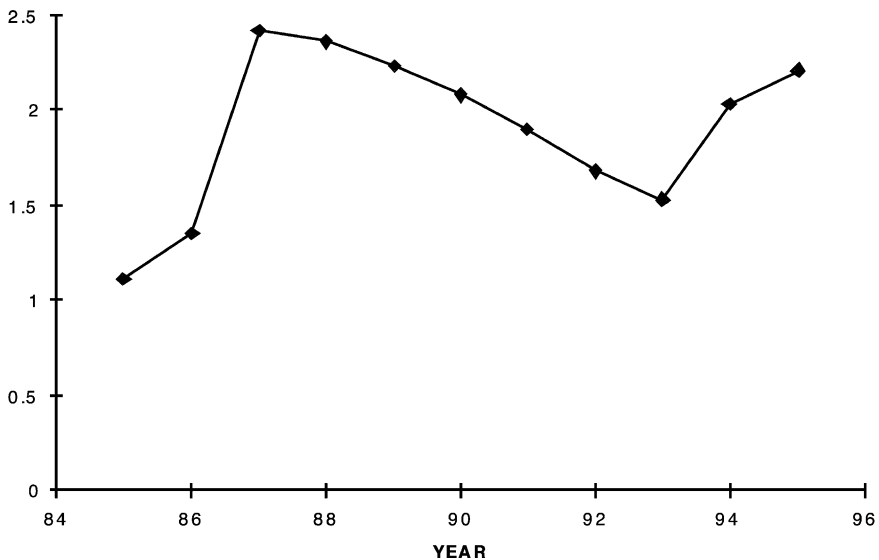


Fig. 3. Time-series of average number of new papers reported in activity reports and listed in working paper series by tenure-track faculty.

post-1991 time period. While the average number of reported papers reaches a local minimum in 1993, the measure declines almost linearly over the 1987–1993 period. Reported papers increased in 1994–1995. New papers listed in the working paper series generally declines from 1985 to 1995. These time series, however, do not control for other variables. We present multiple regression tests in the next section.

5. Findings

Subsection 5.1 begins by testing whether the formation of the CTE was associated with higher teacher ratings (Hypothesis 1). After providing evidence consistent with this hypothesis, Subsection 5.2 tests whether instructors reviewed by the CTE show higher teacher ratings at the time of the evaluation (Hypothesis 2), whether there is a post-adoption time trend in teaching ratings (Hypothesis 3), and whether instructors show higher teaching ratings in courses subsequent to being evaluated by the CTE (Hypothesis 4). Subsection 5.3 discusses the teaching results and describes several sensitivity analyses. Subsection 5.4 presents evidence on whether research output fell (Hypothesis 5).

We adopt a two-step procedure for testing Hypothesis 1 and Hypotheses 2–4 because the formation of the CTE was endogenous during the 1991–1993 period. The same factors motivating the School to start the CTE could also be causing individual instructors being evaluated by the CTE to improve their teacher ratings. To address this endogeneity problem, we test Hypotheses 2–4 using data only following the formation of the CTE (academic years 1994 and later). Following the formation of the CTE, all tenure-track faculty were reviewed and the courses evaluated followed a predetermined schedule (all first year courses followed by all second year courses). Assuming random selection of faculty for evaluation during this period, our analysis tests whether the CTE had an “application” effect.

5.1. *Increased teaching incentives (Hypothesis 1)*

Recall that students rated both the instructor and the course. Thus, there are two measures for each course. We estimate separate regressions for each measure and a third regression of the difference between the instructor and course rating. The Dean’s office and individual faculty focus primarily on the “instructor” rating and not the “course” rating. Students will occasionally report that they do not like the course, but the instructor did a good job. If the formation of the CTE caused faculty to increase teaching effort, and if the instructor rating is more important than the course rating, and if faculty can devote effort at improving their instructor rating more than the course rating, then theory suggests that the difference between the instructor and course rating will increase after the CTE is formed.

We estimate two regression specifications for each of the three dependent variables. One regression includes a separate dummy variable for each faculty member (a fixed-effects model).¹¹ In the second regression, instead of a separate dummy variable for each faculty, we include several faculty characteristics such as full or part-time, tenure or non-tenure track, and so forth. We also include the functional area of the course, for example Accounting or Finance. Both regression specifications also include several control variables: percent of women in the program (%WOMEN), whether the course was an executive MBA class in the U.S. (EDP) or an international course (IEDP), and number of times the course has been taught by the faculty.¹² This last variable (NTAUGHT) controls for faculty learning over time. It is an index that counts the number of times a particular instructor has taught the course in different quarters. For example, suppose an instructor begins teaching a new course by offering one section in the fall, two sections in the winter, and one section in the spring. NTAUGHT for the fall section is 1, the winter sections is 2, and the spring section is 3. A positive coefficient on NTAUGHT indicates either faculty learning over time or a survivor bias (instructors continue to be assigned to those courses where they get higher ratings).

We exclude several right-hand side variables from the regressions because they are endogenous. In particular, class size can both affect teacher ratings (large classes are harder to teach) and can be affected by teaching quality (good instructors attract students).¹³ Besides, a further potential bias exists because we do not have class size, but rather number of student evaluations.

We report OLS results since the coefficients are easy to interpret. The findings are qualitatively identical with a Tobit model or a logistical transformation to account for the fact that teaching ratings are truncated (between 1 and 5). Table 4 contains six regressions. The first three are the fixed-effects regressions with the individual faculty dummy variables (not reported). The last three regressions substitute faculty and course characteristics for the dummy variables. The first regression tests whether *student ratings of the instructor* increase beginning in academic year 1992. The variable POST91 is one for all courses in academic years

¹¹ The faculty dummy variables are arguably endogenous if personnel decisions are affected by teaching performance. Similar results, however, hold if we restrict our analysis to shorter time periods such as 1990–1994 where the composition of faculty remains relatively constant.

¹² As discussed below, the results are robust to the inclusion or exclusion of these and other control variables.

¹³ Mirus (1973) finds a small but statistically significant coefficient of class size on teacher ratings. Calderon et al. (1996, p. 50) survey the extensive literature and report conflicting results of the association between class size and teacher ratings. For completeness, we estimate models which include SIZE as a control variable. Our basic results are unaffected. SIZE is significantly negatively associated with teaching ratings in these tests.

Table 4
Regression tests of Hypothesis 1^a 1983–1997

Variable ^b	Instructor dummy variables ^c			No instructor dummy variables		
	Instr.	Course	Diff.	Instr.	Course	Diff.
INTERCEPT	5.51 (8.43)	3.53 (5.77)	1.92 (5.49)	3.14 (5.50)	3.17 (4.29)	1.26 (3.65)
POST91	0.16 (3.53)	0.15 (3.58)	0.01 (0.55)	0.18 (3.31)	0.17 (3.53)	0.02 (0.78)
ACC				0.22 (4.46)	0.13 (2.85)	0.10 (4.04)
STATS				-0.06 (-1.02)	-0.06 (-1.12)	0.00 (0.05)
BPP				0.24 (3.68)	0.25 (4.43)	-0.01 (-0.39)
CIS				0.18 (3.60)	0.07 (1.59)	0.11 (4.64)
FIN				0.19 (3.90)	0.27 (6.02)	-0.07 (-2.87)
GBA				0.07 (1.19)	-0.07 (-1.51)	0.17 (6.37)
MGC				0.24 (2.99)	-0.04 (-0.56)	0.27 (6.91)
OMG				-0.16 (-2.82)	-0.22 (-4.57)	0.07 (2.72)
ORM				-0.32 (-3.99)	-0.39 (-5.59)	0.08 (1.67)
ECON				0.32 (6.21)	0.26 (5.87)	0.05 (2.17)
FULLTIME				0.18 (4.25)	0.20 (5.39)	-0.02 (-1.04)
TENTRACK				-0.30 (-7.32)	-0.28 (-7.59)	-0.03 (-1.42)
TENDUM				0.05 (1.62)	0.07 (2.51)	-0.02 (-1.00)
YEAR	-0.02 (-2.53)	0.00 (0.10)	-0.02 (-4.33)	-0.01 (-1.08)	0.01 (0.82)	-0.01 (-3.76)
%WOMEN	0.01 (2.18)	0.01 (1.96)	0.00 (0.54)	0.03 (3.30)	0.01 (2.59)	0.01 (1.54)
EDP	-0.21 (-5.52)	-0.13 (-3.80)	-0.09 (-4.42)	-0.03 (-0.77)	0.05 (1.27)	-0.09 (-4.14)
IEDP	-0.19 (-3.93)	-0.09 (-2.02)	-0.11 (-4.50)	-0.08 (-1.78)	0.00 (0.07)	-0.09 (-3.96)
LNTAUGHT	0.10 (5.83)	0.08 (4.90)	0.01 (1.57)	0.11 (7.04)	0.09 (6.48)	0.01 (1.86)
R ²	0.43	0.40	0.30	0.16	0.20	0.12
Obs.	2138	2123	2123	2138	2123	2123

1992 and later. Consistent with Hypothesis 1, POST91’s estimated coefficient is 0.16 with a *t*-statistic of 3.53 after controlling for faculty and other variables. YEAR is a time index. The marginal effect of time is negative. A higher percentage of women in the MBA program increases teacher ratings. Instructors in both domestic and international executive MBA courses receive roughly similar statistically significantly lower ratings. Finally, the logarithm of the number of times a course has been taught (LNTAUGHT) is significantly positive. This suggests that either instructors get better as they teach a class more often or they select out of courses they teach poorly and into courses they teach better. Since the rate of new learning is likely to decrease the more times the course is taught, we use the logarithm of NTAUGHT in the regression.

The second column of Table 4 repeats the same regression except the dependent variable is the *course rating* rather than the instructor rating. The results are similar to those in the first regression—notably POST91 is statistically significant. The third column uses as the dependent variable the difference between the instructor and course ratings. In this model POST91 is statistically insignificant. The insignificance of POST91 in the difference equation indicates that instructors’ additional teaching efforts improved both the students’ rating of the instructor and the course.

Notes to Table 4:

Marketing is the omitted area in the last three regressions which includes area dummy variables.

t-Statistics in parentheses.

^aRegressions of student ratings on a dummy variable (POST91) and control variables to test the hypothesis that establishes the CTE increased course ratings.

^bIndependent variables:

POST91	1 if academic year of the course is greater than 1991, 0 otherwise
ACC	1 if course is Accounting, 0 otherwise
STATS	1 if course is Statistics, 0 otherwise
BPP	1 if course is Business and Public Policy, 0 otherwise
CIS	1 if course is Computers and Information systems, 0 otherwise
FIN	1 if course is Finance, 0 otherwise
GBA	1 if course is General Business Administration, 0 otherwise
MGC	1 if course is Management Communications, 0 otherwise
OMG	1 if course is Operations Management, 0 otherwise
ORM	1 if course is Operations Research, 0 otherwise
ECON	1 if course is Economics, 0 otherwise
FULLTIME	1 if instructor is full-time, 0 otherwise
TENTRACK	1 if instructor is on the tenure track, 0 otherwise
TENDUM	1 if instructor is tenured, 0 otherwise
YEAR	index variable, 1 if course is taught in 1984, 2 if course is taught in 1985,...
%WOMEN	percent of women in the full-time MBA class
EDP	1 if course is an executive development course taught in the U.S. program, 0 otherwise
IEDP	1 if course is an executive development course taught in an international program, 0 otherwise
LNTAUGHT	logarithm of the number of quarters the instructor has taught the course

^cFixed-effects regressions. Includes a separate dummy variable for each instructor (not reported).

POST91 remains statistically significant in the next two regressions where faculty and functional area characteristics are substituted for individual instructor dummy variables. Also of interest, LNTAUGHT remains statistically significant. All of the functional area dummy variables are relative to Marketing. Accounting, Business and Public Policy (BPP), Computers (CIS), Finance, and Economics receive statistically higher ratings than Marketing whereas Operations Management (OMG) and Operations Research (ORM) receive lower instructor and course ratings than Marketing. Full-time faculty (FULLTIME) receive higher ratings than adjunct faculty. However, faculty on the tenure track receive lower ratings than those not on the tenure track. This result can be due to two, not mutually exclusive reasons. Tenure-track faculty are expected to publish whereas non-tenure track faculty are not. Thus, tenure-track faculty substitute effort from teaching tasks to research tasks. A second reason for the result is that non-tenure track faculty are selected and retained primarily for their teaching expertise. TENDUM is one if the instructor is tenured. Its coefficient is positive and marginally statistically significant (t -statistic of 1.62 in the instructor regression and 2.51 in the course regression).

5.2. CTE and longer-run learning / turnover effects (Hypothesis 2–4)

Teaching ratings increase in the academic years following the formation of the CTE. Given this time-series shift in ratings, did the actual evaluation process at the margin increase specific course ratings? Did individual instructors increase their ratings in courses evaluated by the CTE beyond the average secular increase in teacher ratings that occurred after the CTE was formed either in the quarter of evaluation (Hypothesis 2) or subsequent quarters (Hypothesis 4)? Also, was there a positive time trend in teaching ratings in the period after the formation of the CTE due to other learning and turnover effects (Hypothesis 3)?

To test these hypotheses, we make three changes in the Table 4 regressions. First, we restrict the data set to courses taught in academic years 1994 and beyond. By 1994, CTE evaluations were exogenous in the sense that factors leading to the formation of the CTE and to higher teacher ratings had already occurred. Courses chosen to be reviewed by the CTE while not random, followed a well-defined algorithm.

Second, we add two variables to the model, CTE and POSTCTE. CTE takes the value of one if the course in that quarter was evaluated and zero otherwise. In some cases instructors were teaching two sections of the same course being evaluated by the CTE. CTE is equal to one for both sections of the evaluated course.¹⁴ The coefficient on CTE should be positive if being evaluated increases

¹⁴ The results are similar if we average the multiple sections.

teacher ratings. POSTCTE is one if the instructor had previously been evaluated by the CTE. We code all courses taught by the instructor in the post-evaluation period as one because CTE learning effects are potentially transferable to other classes taught by the instructor. The results, however, are similar if we only code as one the same course that was evaluated (for example ACC 401). The coefficient on POSTCTE should be positive if the evaluation process helps the instructor improve teaching as perceived by the students.

Third, we do not include instructor dummy variables because under the maintained hypotheses, faculty composition is endogenous. In assigning faculty to teach particular courses, the Dean's office considers the individual faculty member's past teaching performance. Also, retention and hiring decisions are expected to reflect the increased importance of teaching. Including instructor dummy variables, however, does not alter our basic conclusions about the effects of the CTE.

Table 5 reports the results of the tests. Hypotheses 2 and 4 are not supported. Instructors evaluated by the CTE do not increase student teacher ratings either in the course reviewed or subsequent courses. The coefficients on CTE and POSTCTE are not statistically significant in either regression involving the instructor or course rating. Curiously, POSTCTE is significant in the regression involving the difference between the two ratings. Thus, it appears as if teachers improve their evaluations relative to their courses' evaluations.¹⁵

The results generally support Hypothesis 3. The coefficient on the time trend variable (YEAR) is positive for both the instructor and course ratings. This result is marginally significant for the instructor regression ($p = 0.12$) and significant for the course regression ($p = 0.01$). The difference regression suggests that the gap between the instructor and course ratings narrowed over the 1994–1997 time period.

The sign and significance levels of most of the control variables in Table 5 are similar to their counterparts in Table 4. In particular, the logarithm of the number of times an instructor has taught the course (LNTAUGHT) remains highly significant. Unlike Table 4, FULLTIME, %WOMEN, and EDP are less significant predictors of student ratings.

¹⁵ Note that Table 5 focuses on whether the mean teaching rating differs for those instructors evaluated by the CTE from those not evaluated by the CTE. The CTE process could affect other moments of the distribution rather than the mean (for example, the variance of the ratings). As a more general procedure we use the Kolmogorov–Smirnov test. This test addresses the null hypothesis that the individual teaching ratings of instructors that have been evaluated and those not evaluated are drawn from the same distributions. We are unable to reject the null hypothesis, further supporting our conclusion that the CTE process had no meaningful effect on teaching ratings.

5.3. Discussion

5.3.1. Peer review and teaching ratings

There are several, non-mutually exclusive reasons as to why the CTE evaluation process fails to significantly increase teaching ratings of courses evaluated by the CTE or subsequent courses taught by faculty evaluated by the CTE.

First, the tests lack power. The coefficient on CTE in Table 5 is 0.11, about two-thirds the magnitude as the coefficient on POST91 of 0.16 in Table 4. However, there are only 54 classes evaluated by the Teaching Committee to estimate the CTE coefficient. Thus, its standard error is large. However, the lack of power explanation does not explain the insignificance of the POSTCTE variable. There are about 290 classes taught after the CTE's evaluation.

Second is the lack of direct monetary incentives. Agents are more likely to reallocate effort across tasks if additional incentive compensation is paid for those tasks. In the Simon School, the CTE provides an evaluation report, not additional compensation. The lack of an explicit statement by the Dean's office that compensation would be forthcoming for favorable CTE reports and the subjective performance evaluation system call into question as to how much high-powered

Table 5
Regression tests of Hypotheses 2–4^a 1994–1997

Variable ^b	Instr.	Course	Diff.
INTERCEPT	0.56 (0.29)	-0.64 (-0.36)	2.49 (2.57)
CTE	0.11 (1.23)	0.07 (0.90)	0.03 (0.69)
POSTCTE	0.03 (0.56)	-0.04 (-0.75)	0.06 (2.28)
YEAR	0.03 (1.55)	0.04 (2.52)	-0.02 (-2.59)
ACC	0.24 (2.89)	0.18 (2.54)	0.05 (1.38)
STATS	0.13 (1.43)	0.17 (2.09)	-0.03 (-0.80)
BPP	0.21 (1.91)	0.25 (2.62)	-0.04 (-0.74)
CIS	0.21 (2.55)	0.17 (2.39)	0.05 (1.15)
FIN	0.14 (1.79)	0.24 (3.34)	-0.09 (-2.40)
GBA	0.04 (0.44)	-0.07 (-0.87)	0.16 (3.80)
MGC	0.31 (2.90)	0.03 (0.28)	0.27 (5.25)
OMG	-0.05 (-0.51)	-0.04 (-0.48)	0.01 (0.12)
ORM	-0.26 (-1.09)	-0.37 (-1.75)	0.08 (0.73)
ECON	0.29 (3.44)	0.22 (2.94)	0.07 (1.85)
FULLTIME	0.06 (0.85)	0.12 (2.05)	-0.05 (-1.50)
TENTRACK	-0.26 (-3.68)	-0.21 (-3.31)	-0.07 (-2.15)
TENDUM	-0.02 (-0.43)	-0.01 (-0.16)	-0.00 (0.15)
%WOMEN	0.02 (1.35)	0.01 (0.56)	0.00 (0.59)
EDP	0.04 (0.52)	0.03 (0.52)	-0.01 (-0.19)
IEDP	-0.17 (-2.45)	-0.01 (-0.19)	-0.17 (-4.97)
LNTAUGHT	0.10 (4.30)	0.11 (4.91)	-0.02 (-1.31)
R ²	0.13	0.13	0.20
Obs.	814	799	799

incentives were created by the CTE review. However, CTE reviews still generate peer pressure. Hence, even in the absence of explicit pecuniary compensation provided by the Dean's office, faculty still have incentives to allocate effort to courses evaluated by the CTE.

Third, teacher ratings do not measure "teaching." The use of student ratings, S_i , in Eq. (1) assumes that teacher ratings are positively correlated with "teaching," which is largely unobservable. Even so, student ratings measure "teaching" with error and bias. The CTE was established in part to reduce both measurement error, such as the amount of course development effort, and bias. Note that the variable "CTE_{*i*}" is a dummy variable on whether the *i*th course was evaluated; it is not a measure of teaching. Student ratings can be a biased measure of teaching to the extent ratings are influenced by reduced workloads, giving easy exams, or passing out cookies the last day of class before the evaluations.

Fourth, the CTE process might reduce student satisfaction. Student ratings might fall when the CTE is evaluating the course if the instructor substitutes away from tasks valued by students and towards tasks valued by the CTE. For example, the instructor might increase the course work load and analytic content to impress the CTE but which lowers student ratings. Or, student ratings might fall even

Notes to Table 5:

Marketing is the omitted area in the last three regressions which includes area dummy variables.

t-Statistics are in parentheses.

^aRegressions of student ratings on a dummy variables (CTE and POSTCTE) and control variables to test the hypothesis that CTE course evaluations increased the ratings of the course being evaluated and subsequent offerings of the same course by the instructor.

^bIndependent variables:

CTE	1 if this particular section of the course was evaluated by the CTE, 0 otherwise
POSTCTE	1 if the instructor was previously evaluated by the CTE, 0 otherwise
YEAR	time index variable
ACC	1 if course is Accounting, 0 otherwise
STATS	1 if course is Statistics, 0 otherwise
BPP	1 if course is Business and Public Policy, 0 otherwise
CIS	1 if course is Computers and Information Systems, 0 otherwise
FIN	1 if course is Finance, 0 otherwise
GBA	1 if course is General Business Administration, 0 otherwise
MGC	1 if course is Management Communications, 0 otherwise
OMG	1 if course is Operations Management, 0 otherwise
ORM	1 if course is Operations Research, 0 otherwise
ECON	1 if course is Economics, 0 otherwise
FULLTIME	1 if instructor is fulltime, 0 otherwise
TENTRACK	1 if instructor is on the tenure track, 0 otherwise
TENDUM	1 if instructor is tenured, 0 otherwise
% WOMEN	percent of women in the fulltime MBA class
EDP	1 if course is an executive development course taught in the U.S. program, 0 otherwise
IEDP	1 if course is an executive development course taught in an international program, 0 otherwise
LNTAUGHT	logarithm of the number of quarters the instructor has taught the course

though the instructor is putting more effort into the course being evaluated by the CTE because of the additional psychological stress imposed by the evaluation process. For example, the evaluatee becomes so nervous being observed by other faculty members that lecture delivery suffers.

5.3.2. Organizational or market-driven incentives?

Arguably the labor market for business school professors placed more value on teaching skills after 1992. Thus, faculty might have been expected to invest in teaching capital due to external labor market concerns. In this case, teaching performance might have increased at Rochester even if the changes in internal incentives had little effect.

To address this concern, we obtained the a time series of mean MBA student course ratings from the Graduate School of Business at the University of Chicago. Chicago asks a slightly different question: “How much did you get out of this course?” Fig. 4 plots the mean *course* ratings at both Rochester and Chicago over the sample period, and the mean *instructor* rating at Rochester. Chicago also displayed a significant increase in course ratings from 1982 to 1997 (thinner line). The increase at Chicago is fairly monotonic from 1986, while at Rochester the turning point appears in 1990. The largest increase at Rochester occurred between 1991 and 1992 when incentives were changed at the School. Chicago did not have a particularly large increase between these years.

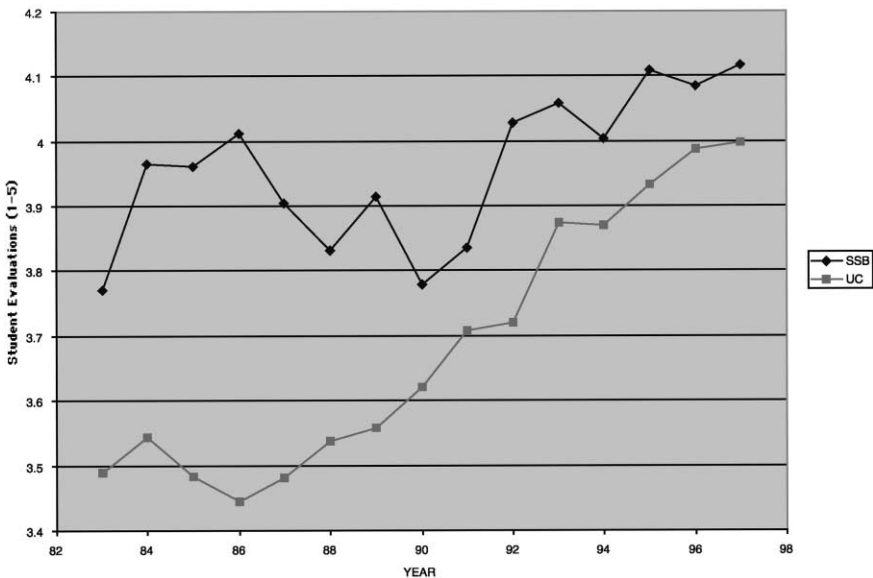


Fig. 4. Universities of Chicago and Rochester: mean student evaluations.

The data in Fig. 4 suggests that teaching ratings may have increased at other top business schools over the sample time period. Such a finding, however, does not necessarily imply that the enhanced teaching performance is market-driven and that organizational incentives are irrelevant. Indeed other schools might have made contemporaneous changes in incentives to enhance teaching. For example, increasing teaching incentives was discussed at MIT at about the same as at the Simon School.

To provide additional evidence on this issue we examine the ratings of adjunct professors at the Simon School. These instructors are typically retired or active business people from the local community. They are not in the national labor market as business-school professors. Nor is there much competition among the local universities for adjunct professors. Given this situation, changes in teaching performance among adjunct professors is more likely to be due to changes in organizational incentives than outside labor market effects. Data previously reported in Table 2A indicates that similar to the full sample, there is a significant increase in teaching ratings for adjunct professors around 1992. Their mean teaching rating in 1990–1991 was 3.78 compared to 4.19 in 1993–1994. The difference is significant at the 0.01 level (t -statistic = 3.73). This finding helps to reinforce our interpretation that the increases in teaching ratings at the Simon School around 1992 were driven at least in part by changes in organizational incentives.

5.3.3. Sensitivity checks

Numerous sensitivity checks assure that our basic results are not driven by omitted variables, choice of functional form, or factors other than the change in organizational incentives. For example, a new classroom building was opened in the Fall of 1991 at the Rochester campus and there were some changes in the evaluation form in the Spring of 1992. Overall, our additional tests suggest that our basic results are robust to a wide range of specification checks. The results also do not appear to be driven by other factors, such as the new building or changes in forms. Here we report some of these sensitivity checks.

1. We restricted the data set to only those instructors who taught before 1991 and were also teaching in 1993 and re-estimated all the models. The results are qualitatively similar to those reported. These results suggest that increased teaching ratings were not due to faculty turnover (replacing teachers having low ratings with teachers having higher ratings).

2. The one control variable that affects the basic results is the percent of international students. Aigner and Thum (1986) find that in undergraduate economic courses, foreign students rated instructors higher than non-foreign students. When we include the percentage of the MBA class that is international students (%INTL), the coefficient on POST91 is no longer statistically significant in any of the regressions in Table 4. The percentage of international students and POST91, however, are highly collinear with a Pearson correlation coefficient of 0.94. The

School increased the international percentage of the MBA class about the same time as it formed the CTE. As is the case with major shifts in corporate strategies whereby numerous policy variables change, both the CTE and the composition of the MBA class shifted. Thus, mean teacher ratings could have increased either due to the formation of the CTE or the larger percentage of foreign students in the program, or both. Fig. 2 provides evidence inconsistent with %INTL causing the increase in student ratings. The time series of teacher ratings in the U.S. executive MBA program follows roughly the same increase in ratings after the formation of the CTE as observed in the entire data set. However, the international composition of the executive MBA did not change significantly over the period. Similarly, the European executive programs experienced a rating increase in the post-1991 period. To provide additional evidence on this issue, we re-estimate the basic regression models using the subsample of executive courses. We find a significantly positive increase in teaching ratings post 1991. These results help assure that the ratings increase for the full sample are not driven by increases in the proportion of international students.

3. The annual return on the New York Stock Exchange for the academic year is included as an additional control variable under the assumption that companies hire MBA students when the stock market is up. And MBA hiring increases general MBA satisfaction including their course ratings. The coefficient on this NYSE variable is usually statistically significant in the instructor and course regressions in Table 4. A 10% return on the NYSE increases teacher ratings 0.20. Including this variable in the model increases the statistical significance of POST91 and reduces the significance of the %WOMEN variable.

4. Using mean teaching rating as our left-hand side variable violates the ordinal nature of the underlying data. Several methods address this problem. First, all the tests were replicated using the median teacher rating which does not violate the ordinal nature of the data. While the significance levels of the coefficient on POST91 fall slightly from those reported in Table 4, none of our inferences change. Second, we estimate an ordered logit and continue to find that the OLS results in Tables 4 and 5 are robust. Third, we create a new left-hand side variable, the percent of the ratings classified as “1” or “2.” Again, the results in Tables 4 and 5 are robust to using this variable.

5. Two other changes occurred about the same time as the CTE formation that could affect our results—a new evaluation form was introduced in Spring 1992 and a new classroom building was opened in Fall 1991. The old form asked 11 questions, the last two asking the questions addressed in this study. The new form ends with the same two questions but before asks three open-ended questions about the strengths and weaknesses of the course and instructor. While we have no predictions regarding how the framing of these questions might affect the results, as a check we re-estimated our basic models eliminating observations from Spring 1992 and beyond. To control for the new building effect, we limit our analysis to classes taught in our overseas executive programs (that were not taught in the new

building). These results are similar to those for the full sample and suggest that the basic findings are not driven by changes in forms or the new building.

6. Another concern is that there was no explicit change in compensation for faculty at the Simon School in 1992. Interestingly, there was an explicit change in compensation for graduate students during the Fall 1992. The School began basing bonuses for teaching labs solely on numerical teaching ratings (from US\$0 to US\$600 depending on student ratings). Teaching ratings for graduate students in the 1989–Summer 1992 subperiod averaged 3.69 vs. 3.91 in the Fall 1992 to Winter 1998 period. These means are significantly different at the 0.01 level indicating an increase in teaching performance after explicit incentives were changed.

5.4. Decreased research incentives (Hypothesis 5)

Table 6 reports the results of regressing both the number of new papers reported on activity reports and listed in the working paper series on POST91, a dummy variable that is one in 1992 and later and zero otherwise and several control variables. Only tenure-track faculty are included in the regression. Two

Table 6
Regression tests of Hypothesis 5,^a 1985–1995

Variable ^b	Reported papers		Listed papers	
	(1)	(2)	(3)	(4)
Faculty Dummies	Yes	No	Yes	No
Intercept	0.474 (0.50)	1.143 * (5.79)	0.036 (0.04)	0.710 * (3.78)
POST91	-0.846 * (-2.61)	-0.001 (-0.01)	-0.159 (-0.50)	-0.617 * (-3.26)
YEARSWORK	0.198 * (2.71)	0.070 (1.51)	-0.031 (-0.43)	0.077 * (1.75)
WORKSQUARE	-0.003 (-1.30)	-0.005 * (-3.15)	-0.003 (-1.09)	-0.006 * (-3.37)
QUANTDUM		1.124 * (5.48)		1.225 * (6.31)
TENDUM		0.998 * (3.65)		0.961 * (3.72)
Observations	387	387	385	385
R ²	0.427	0.125	0.405	0.176

t-Statistics are in parentheses.

^aRegressions of number of new working papers on the dummy variable, POST91, and control variables to test the hypothesis that research output fell when more emphasis was placed on teaching.

^bVariable definitions:

Reported Papers	number of new working papers self-reported in the annual faculty report.
Listed Papers	number of new papers listed in the official working paper series during the year
POST91	1 if academic year of the course is greater than 1991, 0 otherwise
TENDUM	1 if instructor is tenured, 0 otherwise
QUANTDUM	1 if faculty member is Operations Management, Operations Research, or Computers and Information Systems, 0 otherwise
YEARSWORK	number of years on Simon School faculty
WORKSQUARE	number of years on Simon School faculty squared

* Significant at the .01 level.

specifications are estimated, a fixed-effects model with a separate dummy variable for each faculty member and a model with dummy variables for faculty type. If faculty substituted research effort for teaching following the heightened teaching incentives following the creation of the CTE, the coefficient on POST91 is predicted to be negative. Two new independent variables are included: YEAR-SWORK (the number of years on the Simon School faculty) and WORKSQUARE (YEARSWORK^2). We include these variables to capture potential differences in productivity over a faculty member's life cycle. For instance, one might expect research output to increase over the early years of employment, eventually leveling off or decreasing in later years.

The estimated coefficient on POST91 is reliably negative in Table 6 in the reported papers regression with faculty dummies (-0.85) and in the listed papers regression without the faculty dummies (-0.62). POST91 is not reliably negative in the reported papers regression without faculty dummies and in the fixed effects regression using listed papers. Individual faculty vary substantially in their research output as evidenced by the higher R^2 in the fixed effects models (0.427 and 0.405) compared to the R^2 in the regressions with faculty control variables (0.125 and 0.176). In models (1), (2), and (4), faculty with longer work histories (YEARSWORK) produce more papers than faculty with shorter work experience for up to about the first 6 years. Then their research output declines.¹⁶ Tenured faculty (TENDUM) produce about one more working papers per year than untenured faculty. However, this does not control for whether the paper is directed at a refereed journal. For example, along with the increased teaching emphasis the Dean's office encouraged faculty, and especially more senior faculty to publish in popular, professional journals to increase the School's exposure in the business community. Finally, faculty in quantitative areas (Computers, Operations, and Operations research) write about one more working paper per year than faculty in other areas.

The results in Table 6 are robust to several different model specifications (not reported). For example, because the dependent variable is skewed and discrete, we tried various non-linear specifications such as $\log(1 + \text{papers})$, and an ordered-logit regression (coding the dependent variable as "0" if zero papers, "1" if one paper, "2" if two papers, and "3" if three or more papers). The sign and significance of POST91 remains qualitatively the same under these alternative specifications. The results are also relatively robust to the inclusion or exclusion of alternative control variables (e.g., faculty age). If a trend variable (calendar year) is included, the results are reversed for models 2 and 4. POST91 is negative and significant for listed papers, but insignificant for reported papers. If YEARSWORK and WORK-SQUARE are left out of model 3 (where they are insignificant) POST91 is

¹⁶ The negative coefficients for WORKSQUARE, in models (2) and (4) indicate that research output reaches a maximum in 7.0 and 6.4 years, respectively. At the Simon School, promotion to untenured associate occurs in the sixth year and the tenure decision is made around years 8 and 9.

negative and significant at the 0.01 level. Overall, we interpret the results of this analysis as being generally supportive of Hypothesis 5.

6. Conclusions

During the 1990s, there was a substantial environmental shift that increased the importance of teaching relative to academic research at top business schools. The Simon School, like other business schools, changed its performance evaluation and reward systems to increase the emphasis on teaching. One might have expected the effects of these changes to be gradual, given the human capital constraints implied by the composition of existing faculty.

Our results, however, suggest a very rapid adjustment to the changes in incentives. Average teaching ratings increased from about 3.8 to over 4.0 (scale of 5) surrounding the year when incentives were changed. Teaching ratings continue to rise after the changes in incentives, suggesting additional learning and turnover effects. Intense peer review of classes had no obvious effect on teaching ratings for either the evaluated classes or subsequent classes. Finally, we find evidence that suggests faculty-substituted research for teaching following the incentive changes.

Peer review might not be associated with higher student evaluations because of the complementary nature of performance evaluation and compensation (Milgrom and Roberts, 1995). The Dean's office did not formally announce that CTE reviews would explicitly enter the compensation policy of the School. An alternative explanation for the lack of statistical association is that "good" teaching as perceived by faculty evaluators and by students are not highly correlated. For example, the CTE values courses with more intellectual rigor and greater work loads, whereas students tend to value courses with more current business content, more entertaining lectures, and lower work loads. Failure to find a positive association between faculty peer review and student teaching ratings does not necessarily imply that the large opportunity costs of this activity were wasted. Indeed, such an expenditure was potentially important in sending a credible signal to faculty that the Dean's office and senior faculty actually thought that teaching was important to the School.

It is obviously difficult to generalize from a single organization. Nevertheless, the evidence suggests that incentive systems are very powerful, not only in affecting the amount of effort, but also in affecting the allocation of effort across tasks.

Acknowledgements

The John M. Olin Foundation and the Bradley Policy Research Center at the University of Rochester provided financial support for this study. Research

assistance from Daneille Zimmerman and comments from S.P. Kothari, Phil Lederer, Glenn MacDonald, Wilbert Van der Klaauw, Ron Yeaple, Jerry Warner, Ross Watts and seminar participants at the Arizona Finance Symposium, George Washington University, Hong Kong University, National Academy of Sciences Colloquium “Devising Incentives to Promote Human Capital,” and the Universities of Pittsburgh, Rochester, and Texas are gratefully acknowledged. Robert Hamada and Trish Burns graciously provided summary data on course evaluations at the University of Chicago.

References

- Aigner, J., Thum, F., 1986. On student evaluation of teaching ability. *Journal of Economic Education* 17, 243–265 (Fall).
- Banker, R., Lee, S., Potter, G., 1996. A field study of the impact of performance-based incentive plan. *Journal of Accounting and Economics* 21, 195–226 (April).
- Brown, C., 1992. Wage levels and the method of pay. *Rand Journal* 23, 366–375 (Autumn).
- Brickley, J., Smith, C., Zimmerman, J., 2001. *Managerial Economics and Organizational Architecture*, 2nd edn. McGraw-Hill/Irwin, New York.
- Calderon, T., Gabbin, A., Green, B., 1996. A Framework for Encouraging Effective Teaching Report of the Committee on Promoting and Evaluating Effective Teaching American Accounting Association. Center for Research in Accounting Education, James Madison University.
- Feltham, G., Xie, J., 1994. Performance measure congruity and diversity in multi-task principal/agent relations. *Accounting Review* 69, 429–453 (July).
- Gramlich, E., Greenlee, G., 1993. Measuring teaching performance. *Journal of Economic Education* 24, 3–13 (Winter).
- Hemmer, T., 1995. On the interrelation between production technology, structure, and organizational change in manufacturing. *Journal of Accounting and Economics* 19, 209–245 (March–May).
- Holmstrom, B., Milgrom, P., 1991. Multitask principal-agent analysis: incentive contracts, asset ownership, and job design. *Journal of Law, Economics and Organization* 7, 24–52 (Special issue).
- Lazear, E., 1996. Performance Pay and Productivity, NBER Working Paper 5672.
- Martin, J., 1998. Evaluating faculty based on student opinions: problems, implications and recommendations from Deming’s theory of management perspective. *Issues in Accounting Education* 13, 1079–1094 (November).
- Milgrom, P., Roberts, J., 1995. Complementarities and fit: strategy, structure, and organizational change in manufacturing. *Journal of Accounting and Economics* 9, 179–208 (March–May).
- Mirus, R., 1973. Some implications of student evaluation of teachers. *Journal of Economic Education* 5, 35–37 (Fall).
- Wruck, K., Jensen, M., 1994. Science, specific knowledge, and total quality management. *Journal of Accounting and Economics* 18, 247–287 (November).