Non-Response Bias in Student Evaluations of Teaching: The Effect of Student Incentives

Bree J. Lang*
Matthew Lang†

January 2023

ABSTRACT

This paper takes advantage of a policy change at a large public institution that eliminated an incentive to fill out student evaluations of teaching. The policy change resulted in a considerable decrease in response rates, approximately 30 percentage points over three years. We find that the response rate reduction was associated with a 0.046 point increase in mean evaluation score on a 5-point scale. We find weak evidence that non-response bias is larger for male instructors, instructors who identify with Asian racial groups and instructors near the center of the score distribution. There is no evidence that non-response bias is stronger for courses with small enrollments.

Keywords: Non-response bias, Student evaluations of teaching

JEL Codes: I23, J15, J16

*Corresponding Author. 900 University Avenue, University of California, Riverside, Department of Economics, Riverside, CA 92521. Email: blang@ucr.edu.

†University of California, Riverside, Department of Economics. Email: mlang@ucr.edu.

I. Introduction

The potential consequences of student evaluations of teaching (SETs) on major personnel and financial decisions has sparked research across multiple disciplines to investigate whether SETs accurately reflect teaching effectiveness. Studies by Carrell and West (2010) and Braga et al. (2014), which exploit random assignments of students and professors in standardized courses across universities, both find that higher SET scores for instructors are associated with lower grades in subsequent courses. Other research also indicates a positive relationship between grade expectations and SET scores (Isely and Singh, 2005; McPherson, 2006; Babcock, 2010; Butcher et al., 2014). Even under the assumption that SETs can be an accurate measure of quality, there is evidence that SETs are biased against female or minority instructors (Boring, 2017; Mengel et al., 2019).

Another concern with the SET process is that student characteristics are highly correlated with the probability that a SET is submitted (Kherfi, 2011; Goos and Salomons, 2017), which may introduce bias from non-random student participation. To estimate the direction and magnitude of the bias, previous studies use proxies for the probability of participation. The most common proxy is the transition from in-person to online collection of SET responses, which is associated with a large reduction in response rates (Dommeyer et al., 2004; Avery et al., 2006; Burton et al., 2012; Capa-Aydin, 2016; Groen and Herry, 2017). Other studies proxy for response rates with individual absenteeism (Wolbring, 2012; Wolbring and Treischl, 2016) and the declining SET response rates in later terms of the academic year (Goos and Salomons, 2017). While many of these studies utilize student-level data, it is possible that these proxies are correlated with other confounding factors that influence SET scores. This shortcoming, combined with a lack of strong consensus among the studies, means that the effect of non-response bias on SETs is still an open question.

Our study uses an exogenous university-wide policy that removed an incentive for students to fill out evaluations to estimate non-response bias in SETs. Between the spring and fall quarter of 2016, the university changed the software used to upload and post grades. Prior to this policy change, students who did not fill out their evaluations had to wait until a designated

date to see their grades. Students were rewarded for filling out evaluations by receiving their grades as soon as they were posted by an instructor. This incentive was not possible with the new software, so its implementation meant that all students would receive their grades immediately after posting, regardless of whether or not they filled out a SET.

This analysis utilizes a unique panel data set that includes the scores and non-responses from every student, for every class, at a large, public institution between fall 2013 and spring 2019. In the three academic years prior to the incentive removal in the fall 2016 quarter, SET response rates averaged 75%. After the policy change, there was a downward trend in response rates and by the spring quarter of 2019, response rates had fallen to 43% across the campus.

In a two-stage regression framework, we compare the same instructors, teaching the same course, before and after the policy change and use the time that has passed since the implementation of the policy to instrument for response rates. Our results show that there is an upward bias associated with non-response. The reduction in response rates between fall 2016 and spring 2019 is associated with a 0.046 point increase in the mean SET scores evaluating course effectiveness (on a scale of 5).[1] The policy change is also associated with an increase in the fraction of lowest (1 out of 5) and highest (5 out of 5) scores received by faculty members. The fraction of 3's and 4's decreased significantly.

We also explore if non-response bias affects instructors differently by gender, race, pre-policy SET scores and course enrollment. We only find weak evidence suggesting that non-response bias is stronger for male instructors, and instructors with SET scores near the middle of the score distribution prior to the policy change. There is no evidence that courses with smaller enrollments are more likely to be affected by non-response bias. We provide a hypothetical estimate of how the rankings in the post-policy would have changed if every course had achieved a 75% response rate. Assuming that non-response bias affects all courses and faculty the same, there is no meaningful change to

Our first-order regression results are consistent with previous research suggesting that students respond to grade incentives to fill out SETs (Johnson, 2002; Dommeyer et al., 2004;

[1]For context, Boring (2017) finds that male students rate male professors 0.185 to 0.252 points (out of 4) higher than female professors.

Goodman et al., 2015; Lipsey and Shepperd, 2021).[2] An important observation we make is that the reward for filling out SETs does not need to be tied to grades in order to be effective, as students in the current analysis respond strongly to only a delay in receiving access to their grades. Showing the strong response of students to relatively minor participation incentives and the bias associated with non-response is relevant for policy makers in universities that are reevaluating their SET process. The University of Southern California and the University of Oregon have recently overhauled their evaluation of teaching process to rely less on numerical values from SETs (UO, 2022; USC, 2018), but many universities continue to use them in the traditional, numerical form (Becker and Watts, 1999; Becker et al., 2012; Allgood et al., 2015). As more universities strive to improve their teaching evaluation process, it is important that they have an accurate understanding of the consequences of all SET biases that currently exist, including the non-response bias we estimate below.

II. Background

To estimate the role that non-response bias has in the SET process, we take advantage of a unique, student-level data set and an exogenous policy change in the incentive for students to fill out SETs. The institution we examine is a large, public university that is on the quarter system (3 terms per academic year) and has approximately 20,000 undergraduates and 3,000 graduate students in our time period of analysis. Our data covers the fall 2013 quarter to the spring 2019 quarter, which equates to 18 total quarters over six academic years.[3]

In each quarter, we have a record of all students and every course that they enrolled in. For each course that the student takes we are given a unique instructor identifier, the department the course resides in, course enrollment, the grade received by the student, and the SET score that they gave in the course. The evaluation score is blank in the event that a student did not fill out a SET in a course. Students are also given a unique identifier that allows us to

[2]Recent field experiment work shows that nudging students to complete SETs is not effective at increasing student participation (Neckermann et al., 2022).

[3]We omit summer quarters from the analysis, which makes up approximately 10 percent of the sample. Response rates are lower and evaluation scores are higher in summer courses. We estimate and report the results including summer courses in the Appendix and the implications are not meaningfully altered in size or statistical significance.

track their SETs over multiple terms. In the case of both students and instructors, we have information on self-reported gender and race.

The richness of the student-level data is augmented by an exogenous university-wide policy shock that began in the fall 2016 quarter. Before fall 2016, students had an incentive to fill out the SET for a course because doing so allowed them to see their grade in the class as soon as it was uploaded to the registrar by the instructor. If a student did not fill out a SET for a course, they would not be able to see their grade in that course until one week after grades were due. Starting in the fall quarter of 2016, the university changed the central grading software and in the process, the incentive to fill out SETs was eliminated and all students in a course received their official grade at the same time.

The effect of removing the incentive to fill out SETs is seen in figure 1. The dashed line is the mean response rate of SETs across all undergraduate courses in a term. In the three academic years prior to the policy change (2013-14 to 2015-16), the mean response rate is 75 percent and the term-to-term trend is relatively stable. The response rates in the fall quarters are slightly higher than winter and spring quarter response rates.[4] Response rates start to decline in fall 2016. The mean response rate is 58 percent in the three academic years after the policy, 2016-17 to 2018-19, and is 43 percent by the spring quarter of 2019.

Response rates meaningfully decline following the policy change, and it is noteworthy that the effect was not sudden, but instead gradual. The delayed response to the policy is arguably due to the fact that the university did not make an effort to clearly communicate the policy change to students. The original incentive was communicated to students through the quarterly emails they would receive reminding them to fill out SETs. The only way a student would have known the incentive was removed is if they noticed that the language about the incentive had been erased from the quarterly emails. All faculty were alerted to the policy change by email and asked to remind students to fill out SETs.

Figure 1 also reports the average SET score by quarter. There are 19 questions on the SET and we report the results of the question asking students if the "course overall" was an

[4]This consistent decline in response rates over the academic year has been used in prior research to estimate non-response bias in SETs (Goos and Salomons, 2017).

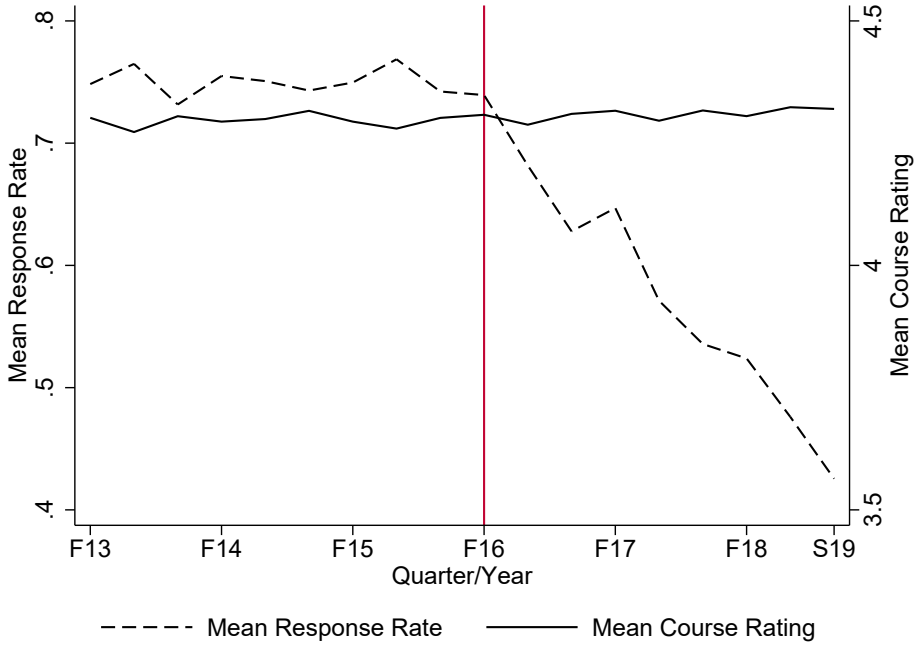

Figure 1: Response Rate and Mean Rating for Course Quality Over Time

Notes: "Mean Course Rating" is the average score received on the evaluation question asking students to respond to the prompt, "This course overall as a learning experience was excellent" on a scale of one (low) to five (high). The incentive to fill out evaluations was removed in fall 2016, which is indicated by the vertical red line. The mean response rate is 0.75 before the policy change, 0.58 after the policy change and 0.43 in the last quarter of observation. The mean evaluation score is 4.30 before the policy change, 4.31 after the policy change in 4.32 in the last quarter of observation.

excellent learning experience. The discrete scoring options range from one (low) to five (high). The stability of the course rating between 2013 and 2019 stands in contrast to the significant reduction in response rates following the policy change. In the years prior to the policy change, the mean score of this question is 4.30 and averages 4.31 in the years after the change. The average is 4.32 in the spring 2019 quarter. Although the aggregate data does not suggest the existence of non-response bias, examining university-level statistics can mask important changes at the course and instructor level. Our data allows us to study the effect of the policy change at a dis-aggregated level in a two-stage regression framework and better identify the potential effect on non-response bias.

One may be concerned that that falling response rates may have been the result of a general decline across all similar universities, as opposed to the change in this specific response incentive at this university. To investigate that possibility, we contacted every university from the same state university system to request average response rates by quarter over this time period. We received data from two universities on response rates from 2013 to 2019. Before the fall quarter of 2016 the average response rate from these universities was 57.2%. After and including the fall quarter of 2016, the average response rate was 56.0%.[5] Although response rates decreased at these universities, the magnitude was quite small relative to the decline we observe in our study.

Before fully utilizing the data in the next section, we note that using scores from other questions in the SET does not meaningfully change outcomes. With the exception of a few questions on the SET asking about student effort, individual SET scores are highly correlated. In the Appendix, we show that the implications of the results do not change when using a different SET question.

[5]One of the comparison universities was transitioning from paper evaluations to online evaluations at this time. If we omit the paper evaluations, the two reported statistics are 58.8% before fall of 2016 and 57.3% after. Online evaluations made up only 8% of evaluations at this institution before fall 2016 and 18% of the evaluations after.

III. Data

As mentioned in the previous section, our data describes SET responses for every student at a large, public university from the fall 2013 quarter to the spring 2019 quarter. We observe if a student fills out a SET for each course, their course grade, self-reported gender and race. Our data can also track courses and instructors over time, observing the instructor's self-reported gender and race.

The raw student-level data is used to construct course-level observations for all undergraduate courses in each of the 18 traditional academic quarters between fall 2013 and spring 2019. For every course, we calculate the mean score of the overall course effectiveness question on the SET, the distribution of scores between one and five, the fraction of students that fill out a SET, the fraction of students in a number of demographic categories and the average grade in the course on a four-point scale.[6]

Table I describes the course-level data for 10,637 courses that were taught between fall 2013 and spring 2019. We drop 537 courses that recorded zero responses (less than five percent), which can occur naturally because an instructor successfully requests to have SET results removed from the record. We also drop 630 observations for singleton observations of courses or instructors (Correia, 2015). Our final data set contains information on 997 instructors who teach an average of 10.7 courses over the 18 quarters we observe. There are 1,486 unique courses taught an average of 7.2 times. There are 711 instructors and 1,176 courses on both sides of the policy change. The average course is taught by 2.1 unique instructors and the average instructor teaches three different courses. Conditional on an instructor teaching the same course both before and after the policy change, they teach the course 2.5 times before and 2.7 times after the incentive was removed.

In Table I, we summarize the data separately for the period when the response incentive was in place (fall 2013 to spring 2016 quarters) and the time period after the incentive was removed (fall 2016 to spring 2019 quarters). The final column reports the p-value of a t-test with the

[6]Students can take courses as pass/fail. We assign passing grades as a C and failing as an F. There are also numerous students that receive "grade delays", we omit these from the average grade calculation unless we find a duplicate record that updates the grade. In cases of multiple grade updates, we keep the highest grade.

Table I: Course Descriptive Statistics

	Before Fall 2016 With Incentive	Fall 2016 & After Without Incentive	p-value
Response Rate	0.75	0.58	<0.001
Mean Evaluation Score	4.30	4.31	0.089
Fraction of Scores			
1	0.017	0.021	<0.001
2	0.031	0.034	0.005
3	0.096	0.089	<0.001
4	0.352	0.325	<0.001
5	0.504	0.530	<0.001
Class GPA	2.82	2.87	<0.001
Upper Division	0.51	0.47	<0.001
Enrollment	0.82	0.78	0.005
Multiple Instructors	0.07	0.06	0.066
STEM Course	0.39	0.38	0.251
Fall Quarter	0.33	0.33	0.571
Winter Quarter	0.35	0.35	0.747
Spring Quarter	0.32	0.33	0.369
Fraction of Students			
Black	0.06	0.05	0.060
Chinese	0.15	0.13	<0.001
Hispanic	0.32	0.36	<0.001
Other	0.14	0.16	<0.001
Other Asian	0.20	0.19	0.012
White	0.13	0.11	<0.001
Not Male	0.51	0.53	<0.001
In Major	0.41	0.38	<0.001
Number of Courses	4,530	6,106	

P-values are for t-tests with the null hypothesis that the mean values are equal before and after the policy change.

null hypothesis that the mean in the two time periods is equal. The average response rate decreased from 0.75 to 0.58 after the policy change, as was mentioned above when discussing Figure 1. The mean score increased marginally between the two periods from 4.30 to 4.31, but the change in the distribution of scores was more pronounced. The fraction of scores that were low (one and two) or very high (five) increased significantly, and scores in the middle of the distribution (three and four) decreased significantly. The finding that the extreme scores increase after the policy aligns with the incentive being more likely to influence students that do not have strong feelings about the course.

There are small, but statistically significant differences in GPA, the fraction of upper division courses and class size before and after the policy change. The average GPA in a course increased from 2.82 to 2.87 after the policy. The fraction of upper division classes decreased from 0.51 to 0.47 over the two time periods and average enrollment decreased from 82.02 to 78.35. The specific quarter of instruction, if the course is classified as STEM and the fraction of students in a particular demographic category are also reported in Table I.[7] We also report when there is more than one instructor in a single course, as students are asked to fill out an evaluation for every instructor of record. We do not have access to student major, but we proxy for this variable with the subject in which each student takes the most courses over the observed time period. In cases of a tie, we assign the subject that comes first alphabetically. Table I indicates that many of these variables are changing over this time period. We expect that response rates and evaluation scores may be related to these variables and are included in the analysis that follows.

In addition to our baseline results, we provide estimates that allow effects to differ for instructors and courses with different characteristics. Table II provides a summary of instructor characteristics we examine: gender, race and the average evaluation score before the policy change. We classify instructors as either male or not male, as there are a small number of instructors that choose a non-binary gender or do not report. The response rates for the two groups are similar, but non-male instructors report higher mean scores. The higher mean score

[7]We proxy for STEM by the college of instruction. Courses in the Colleges of Business, Engineering and Natural Sciences are classified as STEM.

is potentially because non-male instructors are more likely to teach courses in non-STEM fields, which tend to receive higher evaluation scores.[8]

Table II: Description by Instructor or Course Characteristics

		With Incentive		Without Incentive	
Characteristics	N	Response Rate	Score	Response Rate	Score
Gender					
Male	5,698	0.75	4.25	0.57	4.28
Not Male	4,938	0.76	4.35	0.58	4.34
Ethnicity					
White/European	6,632	0.75	4.30	0.57	4.30
URM	1,931	0.75	4.35	0.57	4.37
Asian	2,073	0.76	4.24	0.59	4.27
Pre-Policy Mean Evaluation Quartile					
1st	1,915	0.75	3.87	0.56	3.96
2nd, 3rd	4,965	0.75	4.30	0.57	4.29
4th	2,451	0.75	4.64	0.59	4.61
Unobserved	1,305	–	–	0.58	4.31
Course Enrollment – Median Enrollment = 41					
Below Median	5,380	0.75	4.40	0.59	4.41
Above Median	5,256	0.76	4.20	0.56	4.21

The URM racial category includes instructors who self-report their race as Black Hispanic, Pacific Islander, Native American, other or no response. The Asian racial group includes those who report their race as Chinese or from other Asian countries, also including those from the Middle East or India. Smaller classes enroll 41 or fewer students. Larger classes enroll 42 or more studnets.

Instructors are also classified into groups based on self-reported race. The majority of instructors at the university report their ethnicity as "White/European". We divide the rest of the instructors into two groups. Underrepresented minorities (URM) include Black, Hispanic, Pacific Islander and Native American instructors. We also include instructors that do not report an ethnicity in this group. A second non-white group includes instructors that identify as Chinese (the largest subgroup), being from other Asian countries, India or other Middle Eastern countries. Differences across race may also be the result of instructors in the Asian racial group being relatively more likely to instruct more technical courses.[9]

We also examine effects stratified by the average score an instructor received before the

[8]Non-male instructors teach only 30 percent of STEM courses in our data, compared to 57 percent of other courses. The average evaluation score for STEM courses is 4.37, compared to 4.20 for non-STEM courses.

[9]Of all courses taught by instructors in the Asian racial group, 54 percent are classified as STEM. The comparable statistic for instructors in the white and URM group are 38 and 23 percent, respectively.

policy change. We calculate the average evaluation score across all courses taught by each instructor before the fall 2016 quarter and place each instructor into quartiles. We define 1,305 observations as "Unobserved", meaning that these courses were taught by instructors that did not teach any courses before the policy change. We can only observe these courses after the fall 2016 quarter. The summary statistics in Table II suggest that there are baseline differences among these groups and removing the incentive may have affected each group differently.

The final stratification we examine is based on course size. One may expect that response rate bias will affect courses with lower enrollments more than those with higher enrollments, as the loss of responses may have larger effects on the mean score. We classify courses into two groups, above and below the median class size of 41 students in the sample.[10] The statistics presented at the bottom of Table II show that the average evaluation score is lower for larger classes at the baseline. Removing the incentive reduced the response rate more in larger classes, but the effect on evaluation score is similar between the groups.

IV. Empirical Analysis

A. Specification and Initial Results

The preferred experiment observes the same instructor, teaching the same course (to the same students) before and after the incentive to fill out evaluations is removed. To move closer to that scenario, we utilize the exogenous removal of the incentive to fill out SETs in the first stage of a two-stage regression framework that estimates the relationship between response rates and evaluation scores. Because the policy change affects all observations at the same time, we use a linear time trend to predict response rates instead of quarter/term fixed effects, which is perfectly collinear with the policy change.

In equation 1, a first stage regression estimates how the policy change affects response rates:

$$ResponseRate_{ic} = \alpha_1 AfterPolicy_{ic} + \alpha_2 Trend_{ic} + \alpha_3 AfterPolicyic \times Trend_{ic} + \boldsymbol{X_{ic}} \times \Omega + \kappa_c + \gamma_i + \upsilon_{ic}. \quad (1)$$

[10] Stratifying the sample into more than two groups does not yield additional insights to this comparison.

The variable $ResponseRate_{ic}$ represents the SET response rate in course c, taught by instructor i. We model the policy change using two terms. First, $AfterPolicy_{ic}$ is a binary variable equal to one if the policy change has occurred. It is equal to zero for all quarters before fall 2016 and equal to one thereafter. The variable $Trend_{ic}$ is a linear trend variable centered around the policy change. For example, it is equal to zero in fall 2016, equal to one in winter 2017, and two in spring 2017. It is equal to negative one in spring 2016 and negative two in winter 2016. Including $AfterPolicy_{ic}$ estimates the difference in response rate between the spring 2016 and fall 2016 quarter. The interaction with $Trend_{ic}$ estimates how the response rate trend changes after the policy is enacted in fall 2016. Using a linear trend is consistent with the pattern shown in Figure 1.

The regression includes the control variables from Table I in the matrix $\boldsymbol{X_{ic}}$, as well as course and instructor fixed effects. We estimate the results using robust standard errors because we are estimating for the entire population of courses and the "treatment" is applied uniformly (Abadie et al., 2017). Clustering the standard errors by instructor, course or both does not meaningfully alter the statistical significance of the results.

The results from Equation (1) are reported in the first column of Table III. The F-test statistic for the first stage is equal to 811.81 and both instruments are statistically significant at the one percent level. The estimates indicate that the response rate decreased by 3.47 percentage points in the first quarter after the incentive was removed and an additional 3.70 percentage points each quarter after. The coefficient for the linear trend not interacted with the policy change variable is included in the regression but not statistically significant and aligns with the stable response rates prior to the policy change observed in Figure 1.

Consistent with previous literature, courses with higher grades have higher response rates (Wolbring and Treischl, 2016). Enrollment has a positive, but diminishing effect on response rates. Also consistent with previous literature, the only student demographic variable that is associated with the response rate is the fraction of students that identify as not male (Reisenwitz, 2016). A course with students that are only non-male is estimated to have a 12.8 percentage point higher response rate than a course of only male students. The binary variables for the quarter of instruction indicate that response rates are lower in the winter and

Table III: Effect of Evaluation Incentive Policy Change on Response Rate and Evaluation Scores

	Response	Mean Eval.	Fraction of Each Score Received				
Dependent Variable	Rate	Score	1	2	3	4	5
No Incentive	-0.0346***						
	(0.00455)						
x Trend	-0.0370***						
	(0.000919)						
Predicted Response Rate		-0.143**	-0.0265***	-0.00870	0.0533***	0.168***	-0.186***
		(0.0608)	(0.00840)	(0.00971)	(0.0165)	(0.0276)	(0.0315)
GPA	0.0731***	0.243***	-0.0168***	-0.0189***	-0.0385***	-0.0424***	0.117***
	(0.00527)	(0.0148)	(0.00217)	(0.00251)	(0.00400)	(0.00649)	(0.00712)
Enrollment (100s)	0.0267***	-0.186***	0.00570	0.0146***	0.0390***	0.0415***	-0.101***
	(0.00903)	(0.0251)	(0.00367)	(0.00396)	(0.00530)	(0.00891)	(0.0114)
Enrollment^2 (100s)	-0.00368**	0.0279***	-0.00102*	-0.00227***	-0.00557***	-0.00587***	0.0147***
	(0.00158)	(0.00420)	(0.000562)	(0.000660)	(0.000902)	(0.00149)	(0.00197)
Multiple Instructors	-0.0154	-0.0827**	-0.00224	0.00914	0.0158	0.0326*	-0.0553***
	(0.0163)	(0.0342)	(0.00304)	(0.00557)	(0.0150)	(0.0172)	(0.0185)
Fraction of Students							
Black	-0.0614	0.224**	-0.0238	-0.0156	-0.0366	-0.00819	0.0842
	(0.0393)	(0.0996)	(0.0151)	(0.0158)	(0.0249)	(0.0480)	(0.0532)
Hispanic	-0.0123	0.187***	-0.0142*	-0.0157*	-0.0342**	-0.0148	0.0790**
	(0.0237)	(0.0585)	(0.00777)	(0.00930)	(0.0165)	(0.0296)	(0.0323)
Non-Male	0.128***	-0.0829**	0.00453	0.0116**	-0.00146	0.0328*	-0.0475**
	(0.0147)	(0.0362)	(0.00475)	(0.00575)	(0.0105)	(0.0182)	(0.0203)
Winter Quarter	-0.0118***	0.00918	-0.00206	0.000836	-0.00250	0.00153	0.00218
	(0.00339)	(0.00892)	(0.00135)	(0.00142)	(0.00242)	(0.00381)	(0.00441)
Spring Quarter	-0.0278***	0.0177*	-0.00408***	-0.00367**	-0.000605	0.0109***	-0.00250
	(0.00364)	(0.00933)	(0.00136)	(0.00151)	(0.00257)	(0.00414)	(0.00469)
Stage	First	Second					

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. All regressions include 10,636 observations with instructor and course fixed effects. Although not reported, all of the control variables from Table I are included.

spring quarters, relative to the fall quarter.

The results in column (1) of Table III are used in an instrumental variables approach to predict a response rate for each observation and estimate the effect on the mean SET score.[11] The second stage includes all of the control variables as Specification (1) and is described by the following:

$$SET_{ic} = \beta_1 \widehat{ResponseRate}_{ic} + \beta_2 Trend_{ic} + \boldsymbol{X_{ic}} \times \Phi + \eta_c + \lambda_i + \epsilon_{ic}. \quad (2)$$

In the second column of Table III, the dependent variable in the second stage regression is the mean SET score for a given course and instructor. The coefficient on the predicted response rate is equal to -0.143 and statistically significant at the five percent level. This implies that when the response rate increases by 100 percent, the mean evaluation score decreases by 0.143. In other words, students who are likely to give higher SET scores are more likely to fill out an evaluation. For this specific policy change, this estimate implies that the 32 percentage point reduction in the response rate that occurred from the time of the policy change until the end of the observation period is associated with a 0.046 increase in the mean SET score.

The coefficients for average grade and enrollment align with expectations and previous research. An increase in the average grade in a course by one point on a four-point scale increases the average SET score by 0.243 (McPherson, 2006; Matos-Díaz and Ragan Jr, 2010; Wang and Williamson, 2022). The average SET score decreases as enrollment increases, but at a diminishing rate. Other coefficients indicate that courses with more than one instructor of record receive lower SET scores, higher percentages of Black and Hispanic students are associated with higher SET scores and higher percentages of non-male students are associated with lower SET scores.

To understand how response rates alter the distribution of scores received, we estimate the instrumental variable framework five times with different dependent variables in the second stage. Instead of the mean SET score, the dependent variable is set equal to the fraction of

[11] Coefficients are estimated using the *ivreghdfe* command developed in Correia (2017).

responses that give each possible score of one through five. The results in the last five columns of Table III suggest that as the response rate increases, the fraction of scores that are the lowest or highest decrease, although the magnitude of the decrease in the highest category is larger than in the lowest category. Inducing more students to fill out evaluations increases the fraction of scores in the three and four category. The policy removing the incentive to fill out SETs is associated with an increase in the fraction of ones and fives and a decrease in the fraction of threes and fours. As one might expect, students that do not have strong positive or negative feelings about the course are most affected by a change in the incentive to fill out an SET.

B. Heterogeneous Effects by Instructor and Course Characteristics

We repeat the analysis in the previous subsection but allow for the possibility of heterogenous effects across instructor and class attributes. To construct a gender-specific response rate variable, we interact the response rate with a binary variable equal to one if the instructor identifies as non-male. We then estimate two separate first stage regressions, one for each response rate variable, each including four instruments. The two original instruments are included, a binary variable for the policy change and that variable interacted with a linear trend. Each of these are interacted with a binary variable equal to one if the instructor identifies as non-male and included in the first stage. The first-stage coefficients for the instruments are reported in the first two columns of Appendix Table A.1. All regressions include the same control variables as previously defined.

The two first stage regressions are used to predict response rates for male and non-male instructors separately, which are included in the second stage regression that has mean evaluation score as the dependent variable. The first column of Table IV reports the estimates for the predicted response rate variables. The coefficient on the predicted response rate variable is -0.198 and statistically significant at the five percent level. The coefficient on the interaction between the predicted response rate and the non-male binary variable is positive and not statistically significant. An F-test with the null hypothesis that the sum of the predicted response variable and the interaction term is equal to zero has a p-value of 0.323. While not

Table IV: Heterogeneous Effects by Instructor or Course Characteristics

	(1)	(2)	(3)	(4)	(5)
Predicted Response Rate	-0.198**	-0.109	-0.178**	-0.170**	-0.128**
	(0.0801)	(0.0710)	(0.0777)	(0.0773)	(0.0620)
x Not Male Instructors	0.107				
	(0.122)				
x URM Instructors		-0.000718			
		(0.190)			
x Asian Instructors		-0.145			
		(0.168)			
x Lowest Pre-period Quartile			0.275	0.263	
			(0.182)	(0.181)	
x Highest Pre-period Quartile			0.0694	0.0307	
			(0.141)	(0.140)	
x Unobserved in Pre-period			-0.325**		
			(0.135)		
x High Enrollment					-0.0349
					(0.0399)
Observations	10,636	10,636	10,636	9,269	10,636

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. All regressions include instructor and course fixed effects. Although not reported, all of the control variables from Table I are included. High enrollment courses defined as those with 42 or more students enrolled.

conclusive, the results suggest that non-response bias may be stronger for male instructors than female instructors.

We do a similar analysis for instructors of different races. As described in the Data section, we group instructors into those who identify as white/European, those who identify as Black, Hispanic, or other underrepresented groups, and those who identify as Chinese or from other Asian countries, including India and Pakistan. The first stage results are reported in Appendix Table A.1 and the second stage results are reported in column (2) of Table IV. The predicted response specific to the white/European group is negative but not statistically significant. The interacted response rate coefficients are similarly insignificant, however an F-test with the null hypothesis that the sum of the predicted response variable and the interaction term with Asian instructors is equal to zero has a p-value of 0.09. These results suggest little difference in non-response bias by race of instructor, but the largest effect may be for instructors in the group

we classify as Asian.[12]

Next, we separate instructors into quartiles based on their average score in the period before the incentive was changed. We run two instrumental variable regressions for this analysis. One that includes the instructors that were not observed in the pre-period in their own group and one that omits that group entirely. The first stage results for the regression that includes the entire sample is reported in Appendix Table A.2. Because the unobserved group is not observed in the pre-period, we can only include one instrument for that group in the first stage.

The second stage results using the entire sample are reported in the third column of Table IV. The results in that column suggest that non-response bias is strongest for instructors with average scores in the pre-period interquartile range and the instructors who were not observed during the pre-period. The coefficient specific to the interquartile range is equal to -0.178 and statistically significant at the five percent level. The interaction term with the unobserved group is -0.325 and also statistically significant at the five percent level. The larger coefficient may be the result of relatively large improvements to the evaluation scores of new instructors over their first quarters of teaching.

Neither interaction term for the lowest and highest pre-period quartiles are statistically significant. The interaction term for the lowest quartile is equal to relatively large and positive, equal to 0.275. An F-test that the sum of the primary coefficient and interaction term is equal to zero has a p-value of 0.55. For instructors in the highest quartile, the interaction term is equal to 0.0694 and a similar F-test has a p-value of 0.36. The results suggest that instructors with scores near the bottom of the distribution are least likely to be affected by non-response bias. In column (4), we remove the instructors that are not observed before the policy change and it does not meaningfully change the results, relative to column (3).

Finally, we estimate effects separately for courses with enrollment below and above the median. The first stage results are reported in Appendix Table A.2 and the second stage results are reported in column (5) of Table IV, where the omitted group includes courses with enrollments of 41 students or less. The coefficient for the omitted group is equal to -0.128

[12]In the gender-based analysis, we classify those who do not report as male. In the race-based analysis, we classify those that do not report as underrepresented. The results of the analysis do not meaningfully change if those instructors are dropped from the analysis.

and statistically significant at the five percent level. The interaction term for high enrollment courses is equal to -0.0399 and not statistically different from zero. There is no evidence that smaller classes are more likely to be affected by non-response bias.

We provide several tests of robustness in the Appendix. Table A.3 compares the coefficients using different questions on the SET. The first column briefly describes the SET question used as the dependent variable in each row of the table. The final row of the table provides the coefficient from the primary analysis for comparison. The range of coefficients in separate regressions using every other question on the SET is between -0.128 and -0.343. Every coefficient is statistically significant at the five percent level or lower. The two highest coefficients in magnitude are associated with questions about course attendance (-0.306) and effort (-0.343). These relatively large coefficients are consistent with the least engaged students being less likely to fill out an SET without an incentive. Although the differences are small, the coefficients tend to be larger in magnitude for questions about the instructor, relative to questions about the course.

We also provide the results in reduced form in the top panel of Appendix Table A.4. The reduced form estimates are consistent with the primary analysis and confirm that changes in the SET scores are driven by a gradual reduction in response rates. To explore this further, we re-estimate the reduced form results in the bottom panel of Table A.4 without the linear trend interaction term.[13] We still observe non-response bias without the linear trend, although the coefficients are smaller in magnitude and weaker in statistical significance. A similar outcome is seen when we estimate a two-stage regression without the linear trend, as seen in Appendix Table A.5.

Given the linear downtrend in response rates is an important part of our analysis, we examine how many years of data after the policy change are needed to find statistically significant results. That analysis is presented in Appendix Table A.6. As shown in the top panel of that table, removing the final year of observation does not meaningfully affect the first stage results. In the second stage, the magnitude of the coefficient of interest is similar to the results using the entire sample, but it is no longer statistically significant. The results for the distribution of

[13]This regression omits the linear trend variable entirely, as its inclusion captures the change in response rates even if it is not interacted with the policy change variable.

scores, however, has similar magnitudes and statistical significance. Although the signs of the coefficients remain the same, there is little statistical evidence of non-response bias when we remove the last two years of observations in the bottom panel of Appendix Table A.6. This is not surprising, given the reduction in response rates is only seven percentage points (0.75 to 0.68) in the first year after the policy change, relative to the pre-period.

Finally, we re-estimate our baseline results including the summer quarters in Appendix Table A.7 and stratify the results by STEM classification in Appendix Table A.8. Including the summer quarters does not affect the results and we do not find evidence that non-response bias differs between courses classified as STEM or non-STEM.

V. Discussion and Conclusion

The main finding in our analysis above is that the exogenous policy change that removed the incentive to fill out SETs decreased response rates significantly and is associated with a significant increase in the average evaluation score. The policy decreased response rates by approximately 32 percentage points over nine quarters, representing a 44 percent reduction in response rates from the pre-policy period. The policy is associated with a 0.046 increase in the average SET score, which is approximately 0.10 of a standard deviation increase in the mean score.

We find that the effects of the policy change differs for instructors based on their position in the distribution of average scores before the policy change. The largest positive effects from decreasing response rates are found for instructors near the middle of the pre-period distribution. To provide some context for the relative benefit to these instructors, we use estimated effects for instructors in each quartile to conduct a back-of-the-envelope calculation of how the instructor rankings changed as a result of the policy change. The thought experiment asks how average evaluation scores would differ after the policy change if response rates for every course was equal to 0.75. We calculate the hypothetical score for every course and then take the average of the true score and the hypothetical score by instructor. We omit instructors that were not

observed in the pre-period.[14]

The back-of-the-envelope calculation indicates that the ranking of average instructor in the inter-quartile range increased by eight as a results of decreasing response rates. This ranking increase comes at the expense of those in the lowest quartile, where the average ranking decreased by 15. There was no change to the ranking for instructors in the highest quartile. This calculation highlights that our results suggest non-response bias will have the most meaningful effects for those on the margin of the bottom of the score distribution. Lower response rates are least likely to increase the scores of these instructors, so they appear as stronger outliers when response rates are low.

The magnitude of our findings are comparable to previous work examining non-response bias and studies of SET bias more broadly. Dommeyer et al. (2004) and Avery et al. (2006) compare how response rates and SET scores change when professors use online evaluations in one section and in-class evaluations in another. They find that response rates decrease considerably when SETs are completed online, but there is no a significant change in average scores. Using a Heckman selection model, Nowell et al. (2014) also takes advantage of the difference in response rates between in-class and online SETs, but does not find that the selection bias is associated with a meaningful change in average scores.

Capa-Aydin (2016) finds that online evaluations are filled out less often than in-class evaluations (31% to 40%) and in contrast to our paper, the average score for in-class SETs was significantly higher than online. Similarly, work by Mitchell and Morales (2018) show that mandatory online evaluations decreased response rates in on-campus criminology courses from 70% to 53%. The corresponding mean SET scores decreased from 4.51 to 4.39. Layne et al. (1999) find bias in the opposite direction, showing online respondents were significantly more likely to be satisfied with the course than in-person respondents. A drawback of these studies is that estimating non-response bias using the change from in-person to online SETs can influence average scores through channels other than response rates. Our research design is fundamentally different because we take advantage of changes in response rates from an

[14]If included, the calculations indicate that this group actually benefits the most from the policy change at the expense of all other groups. It may be reasonable to assume that the majority of these instructors would fall into the inter-quartile range.

exogenous policy shock, holding the method of SET constant.

Wolbring and Treischl (2016) take a different approach and use multiple evaluations throughout a course to show that absenteeism is higher in the later SET of the term, but the average rating is not changed. It is not clear if the conclusions about non-response bias transfer to other settings, since the study took place in 27 classes during a single summer term at one university. Research by Goos and Salomons (2017) estimates non-response bias by taking advantage of the consistent lower response rate in second-semester courses relative to first-semester courses. They estimate that the existence of non-response bias increases evaluations scores by 0.1332, equivalent to 28% of a standard deviation in evaluation scores. Their estimates are notably larger than ours, but we are not able to eliminate non-response bias that existed prior to the policy change.

Our 0.10 standard deviation increase in SET scores from the policy is also comparable to the magnitude of gender bias found in SETs by Boring (2017) and Mengel et al. (2019). Specifically, Boring (2017) shows that female students rate female professors up to 0.110 points (out of four) lower than male professors. Mengel et al. (2019) use a random allocation of students to show that male students rate female instructors 20.7% of a standard deviation lower than male instructors.

The magnitude of our results can also be compared to research examining the relationship between grades and SET scores. Isely and Singh (2005) estimate that a one unit increase in grade expectation is associated with one standard deviation improvement in SETs scores. Using an exogenous change to grade requirements, Butcher et al. (2014) show that the policy reduced course GPAs by one-sixth of a letter grade on average, and average SET scores decreased by 0.11 on a scale of 4.[15]

Although the current paper focuses on the relationship between response rates, non-response bias and SET scores, a secondary contribution is evidence that students respond strongly to incentives to fill out evaluations. Neckermann et al. (2022) use a randomized field experiment to show that nudging students to participate in SETs does not work. Nudging has been shown to be effective in certain scenarios, such as energy conservation (Allcott, 2011; Allcott and

[15]Standard deviations are not reported in Butcher et al. (2014).

Rogers, 2014), but according to Neckermann et al. (2022) and our first-stage regressions, many students need a personal incentive to fill out evaluations. Butcher et al. (2014) support this claim in noting that in their quasi-experimental setting, response rates are nearly 100% because "students submit evaluations electronically and their ability to see grades in a timely fashion is tied to submitting an evaluation during a specified period."

A primary contribution of our analysis is that we are able to combine rich student-level data with an exogenous policy change that significantly reduced response rates. These results are useful for universities looking to augment or overhaul their current SET process. While we are confident in our identification strategy, we refrain from claiming a causal relationship between response rates and SET scores. Given the new policy was instituted across the entire university, our analysis does not have a typical counterfactual for comparison, so we must make assumptions that the pre-period trends would have continued in absence of the policy. Furthermore, student composition is not held constant across classes and the majority of students did not react to the policy immediately. While these concerns do not negate our findings, it is possible that they affect the estimated magnitudes we report. Nonetheless, the strong relationship we find between the exogenous policy change, response rates and evaluation scores is important to consider when using SET scores to make important financial and personnel decisions.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Allcott, H. (2011). Social norms and energy conservation. *Journal of public Economics 95*(9-10), 1082–1095.

Allcott, H. and T. Rogers (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review 104*(10), 3003–37.

Allgood, S., W. B. Walstad, and J. J. Siegfried (2015). Research on teaching economics to undergraduates. *Journal of Economic Literature 53*(2), 285–325.

Avery, R. J., W. K. Bryant, A. Mathios, H. Kang, and D. Bell (2006). Electronic course

evaluations: does an online delivery system influence student evaluations? *The Journal of Economic Education 37*(1), 21–37.

Babcock, P. (2010). Real costs of nominal grade inflation? new evidence from student course evaluations. *Economic inquiry 48*(4), 983–996.

Becker, W. E., W. Bosshardt, and M. Watts (2012). How departments of economics evaluate teaching. *The Journal of Economic Education 43*(3), 325–333.

Becker, W. E. and M. Watts (1999). How departments of economics evaluate teaching. *American Economic Review 89*(2), 344–349.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics 145*, 27–41.

Braga, M., M. Paccagnella, and M. Pellizzari (2014). Evaluating students' evaluations of professors. *Economics of Education Review 41*, 71–88.

Burton, W. B., A. Civitano, and P. Steiner-Grossman (2012). Online versus paper evaluations: differences in both quantitative and qualitative data. *Journal of Computing in Higher Education 24*(1), 58–69.

Butcher, K. F., P. J. McEwan, and A. Weerapana (2014). The effects of an anti-grade-inflation policy at wellesley college. *Journal of Economic Perspectives 28*(3), 189–204.

Capa-Aydin, Y. (2016). Student evaluation of instruction: comparison between in-class and online methods. *Assessment & Evaluation in Higher Education 41*(1), 112–126.

Carrell, S. E. and J. E. West (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy 118*(3), 409–432.

Correia, S. (2015). Singletons, cluster-robust standard errors and fixed effects: A bad mix. *Technical Note, Duke University*.

Correia, S. (2017). Linear models with high-dimensional fixed effects: An efficient and feasible estimator. Technical report. Working Paper.

Dommeyer, C. J., P. Baum, R. W. Hanna, and K. S. Chapman (2004). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education 29*(5), 611–623.

Goodman, J., R. Anson, and M. Belcheir (2015). The effect of incentives and other instructor-driven strategies to increase online student evaluation response rates. *Assessment & Evaluation in Higher Education 40*(7), 958–970.

Goos, M. and A. Salomons (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Research in Higher Education 58*(4), 341–364.

Groen, J. F. and Y. Herry (2017). The online evaluation of courses: Impact on participation rates and evaluation scores. *Canadian Journal of Higher Education 47*(2), 106–120.

Isely, P. and H. Singh (2005). Do higher grades lead to favorable student evaluations? *The Journal of Economic Education 36*(1), 29–42.

Johnson, T. (2002). Online student ratings: Will students respond? *Online student ratings of instruction: New directions for teaching and learning* (96), 49–59.

Kherfi, S. (2011). Whose opinion is it anyway? determinants of participation in student evaluation of teaching. *Journal of Economic Education 42*(1), 19–30.

Layne, B. H., J. R. DeCristoforo, and D. McGinty (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education 40*(2), 221–232.

Lipsey, N. and J. Shepperd (2021). Examining strategies to increase student evaluation of teaching completion rates. *Assessment & Evaluation in Higher Education 46*(3), 424–437.

Matos-Díaz, H. and J. F. Ragan Jr (2010). Do student evaluations of teaching depend on the distribution of expected grade? *Education Economics 18*(3), 317–330.

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *The Journal of Economic Education 37*(1), 3–20.

Mengel, F., J. Sauermann, and U. Zölitz (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association 17*(2), 535–566.

Mitchell, O. and M. Morales (2018). The effect of switching to mandatory online course assessments on response rates and course ratings. *Assessment & Evaluation in Higher Education 43*(4), 629–639.

Neckermann, S., U. Turmunkh, D. van Dolder, and T. V. Wang (2022). Nudging student participation in online evaluations of teaching: Evidence from a field experiment. *European Economic Review 141*, 104001.

Nowell, C., L. R. Gale, and J. Kerkvliet (2014). Non-response bias in student evaluations of teaching. *International Review of Economics Education 17*, 30–38.

Reisenwitz, T. H. (2016). Student evaluation of teaching: an investigation of nonresponse bias in an online context. *Journal of marketing education 38*(1), 7–17.

UO (2022). Revising university of oregon's teaching evaluations. Accessed December 16, 2022.

USC (2018). University of southern california teaching evaluations update. Accessed December 16, 2022.

Wang, G. and A. Williamson (2022). Course evaluation scores: valid measures for teaching effectiveness or rewards for lenient grading? *Teaching in Higher Education 27*(3), 297–318.

Wolbring, T. (2012). Class attendance and students' evaluations of teaching: Do no-shows bias course ratings and rankings? *Evaluation Review 36*(1), 72–96.

Wolbring, T. and E. Treischl (2016). Selection bias in students' evaluation of teaching. *Research in Higher Education 57*(1), 51–71.

VI. Appendix

Table A.1: First Stage Heterogeneous Effects by Instructor Gender and Race

	(1)	(2)	(3)	(4)	(5)
Response Dependent Var.	All	Non-Male	All	URM	Asian
No Incentive	-0.0398***	-0.00577***	-0.0348***	-0.00137*	-0.000843
	(0.00619)	(0.00133)	(0.00578)	(0.000706)	(0.000713)
x Trend	-0.0374***	4.71e-05	-0.0388***	0.000110	-8.42e-05
	(0.00125)	(0.000229)	(0.00117)	(0.000121)	(0.000129)
x Not Male	0.0113	-0.0187***			
	(0.00890)	(0.00658)			
x Trend x Not Male	0.000879	-0.0363***			
	(0.00184)	(0.00136)			
x URM Instructor			0.00560	-0.0221**	0.000843
			(0.0117)	(0.0103)	(0.00126)
x Trend x URM			0.00435*	-0.0347***	8.23e-05
			(0.00247)	(0.00221)	(0.000256)
x Asian Instructor			-0.00284	0.00225*	-0.0363***
			(0.0115)	(0.00121)	(0.00988)
x Trend x Asian			0.00544**	-0.000188	-0.0328***
			(0.00238)	(0.000218)	(0.00206)

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. All regressions include 10,636 observations with instructor and course fixed effects. Although not reported, all of the control variables from Table I are included. Omitted category is male in columns (1) and (2) and white/European in columns (3)-(5).

Table A.2: First Stage Heterogeneous Effects by Instructor Pre-period Rating Quartile and Course Enrollment

	(1)	(2)	(3)	(4)	(5)	(6)
Response Dependent Var.	All	Q1	Q4	Unobs.	All	High Enrollment
No Incentive	-0.0411***	-0.00210**	-0.000986	-0.00275***	-0.0360***	-0.202***
	(0.00620)	(0.000815)	(0.000792)	(0.000668)	(0.00657)	(0.00687)
x Trend	-0.0397***	-0.000354**	-6.70e-05	-8.15e-06	-0.0382***	-0.0335***
	(0.00129)	(0.000139)	(0.000128)	(0.000109)	(0.00137)	(0.00119)
x Pre-period Lowest Quartile	0.00604	-0.0261***	0.000772	0.000892		
	(0.0115)	(0.00980)	(0.00110)	(0.00114)		
x Trend x Q1	0.000101	-0.0395***	0.000168	4.36e-05		
	(0.00241)	(0.00202)	(0.000221)	(0.000250)		
x Pre-period Highest Quartile	0.0224**	0.00134	-0.0142	0.000968		
	(0.0111)	(0.000991)	(0.00932)	(0.00103)		
x Trend x Q4	0.00973***	0.000514**	-0.0304***	8.66e-05		
	(0.00233)	(0.000204)	(0.00195)	(0.000183)		
x Trend x Pre-period Unobserved	0.00137	6.06e-05	-7.28e-05	-0.0367***		
	(0.00190)	(0.000237)	(0.000203)	(0.00164)		
x High Enrollment					0.00139	0.390***
					(0.00725)	(0.00958)
x Trend x High Enrollment					0.00211	0.0341***
					(0.00156)	(0.00181)

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. All regressions include 10,636 observations with instructor and course fixed effects. Although not reported, all of the control variables from Table I are included. Omitted category is the interquartile range for columns (1)-(4) and low enrollment courses for columns (5)-(6). Note that an indicator variable for instructors who are not observed in the pre-period is collinear with the "no incentive" variable, so it is omitted. High enrollment courses are those with 42 or more students enrolled.

Table A.3: Second Stage Results for All SET Questions

	Predicted Response Rate	
Question	Coefficient	Std. Error
1. Student desire to take course	-0.143**	(0.0608)
2. Student attended class regularly	-0.305***	(0.0430)
3. Student effort invested	-0.342***	(0.0420)
4. Student gained understanding	-0.184***	(0.0519)
5. Time outside of class	-0.128**	(0.0589)
6. Inst. prepared/organized	-0.291***	(0.0538)
7. Inst. used time effectively	-0.264***	(0.0573)
8. Inst. clear/understandable	-0.296***	(0.0610)
9. Inst. enthusiastic	-0.259***	(0.0478)
10. Inst. respect students	-0.228***	(0.0554)
11. Inst. available/helpful	-0.201***	(0.0520)
12. Inst. fair	-0.235***	(0.0583)
13. Inst. effective overall	-0.227***	(0.0594)
14. Syllabus clear	-0.283***	(0.0498)
15. Exams reflect material	-0.235***	(0.0523)
16. Readings contributed	-0.133**	(0.0525)
17. Assignments contributed	-0.169***	(0.0516)
18. Supplements informative	-0.229***	(0.0527)
19. Course overall excellent	-0.143**	(0.0608)

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. All regressions include 10,636 observations with instructor and course fixed effects. Although not reported, all of the control variables from Table I are included.

Table A.4: Reduced Form Results

	Mean Eval.	Fraction of Each Score Received				
	Score	1	2	3	4	5
Panel A: Including Linear Interaction Term						
No Incentive	-0.0120	0.00182	0.000570	-0.00123	0.00546	-0.00661
	(0.0112)	(0.00147)	(0.00180)	(0.00317)	(0.00507)	(0.00575)
x Trend	0.00526**	0.000982***	0.000322	-0.00197***	-0.00622***	0.00688***
	(0.00224)	(0.000310)	(0.000359)	(0.000611)	(0.00101)	(0.00115)
Panel B: Not Including Linear Interaction Term						
No Incentive	-0.0106*	0.00392***	0.00336***	-0.00150	-0.0122***	0.00638*
	(0.00635)	(0.000866)	(0.00105)	(0.00178)	(0.00298)	(0.00331)

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. All regressions include 10,636 observations with instructor and course fixed effects. Although not reported, all of the control variables from Table I are included.

Table A.5: Instrumental Variable Results Excluding the Linear Trend its Interaction

	Response	Mean Eval.	Fraction of Each Score Received				
Dependent Variable	Rate	Score	1	2	3	4	5
No Incentive	-0.173***						
	(0.00312)						
Predicted Response Rate		0.0614*	-0.0227***	-0.0194***	0.00868	0.0705***	-0.0370*
		(0.0368)	(0.00502)	(0.00606)	(0.0103)	(0.0172)	(0.0192)
Stage	First	Second					

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. All regressions include 10,636 observations with instructor and course fixed effects. Although not reported, all of the control variables from Table I are included.

Table A.6: Results Dropping Years at End of Observation Period

	Response	Mean Eval.	Fraction of Each Score Received				
Dependent Variable	Rate	Score	1	2	3	4	5
Panel A: Removing Final Year of Observation							
No Incentive	-0.0325***						
	(0.00491)						
x Trend	-0.0370***						
	(0.00133)						
Predicted Response Rate		-0.133	-0.0315***	-0.00207	0.0733***	0.119***	-0.159***
		(0.0831)	(0.0113)	(0.0128)	(0.0225)	(0.0380)	(0.0430)
Observations				8,233			
Panel B: Removing Final Two Years of Observation							
No Incentive	-0.0216***						
	(0.00620)						
x Trend	-0.0493***						
	(0.00403)						
Predicted Response Rate		-0.0287	-0.0215	-0.0206	0.0626*	0.0513	-0.0718
		(0.132)	(0.0171)	(0.0218)	(0.0365)	(0.0607)	(0.0683)
Observations				6,077			
Stage	First			Second			

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. Although not reported, all of the control variables from Table I are included.

Table A.7: Results Including Summer Quarters

	Response	Mean Eval.	Fraction of Each Score Received				
Dependent Variable	Rate	Score	1	2	3	4	5
No Incentive	-0.0404***						
	(0.00442)						
x Trend	-0.0273***						
	(0.000681)						
Predicted Response Rate		-0.150**	-0.0276***	-0.0138	0.0596***	0.183***	-0.201***
		(0.0599)	(0.00841)	(0.00980)	(0.0165)	(0.0278)	(0.0315)
Winter Quarter	-0.0168***	0.00895	-0.00250*	0.000677	-0.00186	0.00275	0.000937
	(0.00334)	(0.00878)	(0.00134)	(0.00140)	(0.00236)	(0.00378)	(0.00435)
Spring Quarter	-0.0390***	0.0151	-0.00434***	-0.00381**	0.000579	0.0125***	-0.00490
	(0.00357)	(0.00924)	(0.00137)	(0.00151)	(0.00253)	(0.00416)	(0.00468)
Summer Quarter	-0.140***	0.115***	-0.00826**	-0.0149***	-0.0123	-0.0127	0.0481***
	(0.0151)	(0.0329)	(0.00380)	(0.00525)	(0.0116)	(0.0163)	(0.0183)
Stage	First	Second					

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. All regressions include 11,817 observations with instructor and course fixed effects. Although not reported, all of the control variables from Table I are included.

Table A.8: Second Stage Results by STEM Classification

	Mean Eval.	Fraction of Each Score Received				
	Score	1	2	3	4	5
Pred. Response Rate	-0.142**	-0.0302***	-0.00680	0.0495***	0.184***	-0.197***
	(0.0678)	(0.00969)	(0.0108)	(0.0186)	(0.0317)	(0.0357)
x STEM	-0.0150	0.0135	-0.00314	0.0111	-0.0517*	0.0303
	(0.0625)	(0.00867)	(0.0103)	(0.0164)	(0.0278)	(0.0318)

***p<0.01, **p<0.05, *p<0.1. Robust standard errors are reported in the parentheses. All regressions include 10,636 observations with instructor and course fixed effects. Although not reported, all of the control variables from Table I are included. STEM courses are those in the Colleges of Business, Engineering and Natural Sciences.