November 2021

iRisk WORKING PAPER SERIES

2021-iRisk-03

Nudging Student Participation in Online Evaluations of Teaching: Evidence from a Field Experiment

Susanne Neckermann
Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, USA,
(sneckermann@uchicago.edu)

Uyanga Turmunkh
Department of Economics and Quantitative Methods, IESEG School of Management, 3 Rue de la Digue, 59000 Lille, France
LEM-CNRS 922
(u.turmunkh@ieseg.fr)

Dennie van Dolder
School of Business and Economics, Vrije Universiteit (VU) Amsterdam, Gustav Mahlerplein 117, 1082 MS Amsterdam, the Netherlands
Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom
(d.van.dolder@vu.nl)

Tong V. Wang
Institute for Advanced Economic Research, Dongbei University of Finance and Economics, 217 Jianshan Street, Dalian, Liaoning 116025, P.R. China
(t.wang@dufe.edu.cn)

IÉSEG School of Management Lille Catholic University 3, rue de la Digue F-59000 Lille Tel: 33(0)3 20 54 58 92
www.ieseg.fr

Nudging Student Participation in Online Evaluations of Teaching: Evidence from a Field Experiment

Susanne Neckermann[a], Uyanga Turmunkh (corresponding author)[b,c], Dennie van Dolder[d,e], Tong V. Wang[f]

Forthcoming in the European Economic Review

[a] Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, USA, sneckermann@uchicago.edu; [b] Department of Economics and Quantitative Methods, IESEG School of Management, 3 Rue de la Digue, 59000 Lille, France, u.turmunkh@ieseg.fr; [c] LEM-CNRS 9221; [d] School of Business and Economics, Vrije Universiteit (VU) Amsterdam, Gustav Mahlerplein 117, 1082 MS Amsterdam, the Netherlands, d.van.dolder@vu.nl; [e] Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom, d.vandolder@essex.ac.uk; [f] Institute for Advanced Economic Research, Dongbei University of Finance and Economics, 217 Jianshan Street, Dalian, Liaoning 116025, P.R. China, t.wang@dufe.edu.cn

Abstract. This paper reports the results of a large randomized field experiment that investigates the extent to which nudges can stimulate student participation in teaching evaluations. The three nudges that we used were designed to either: (1) heighten students' perceived impact of teaching evaluations, (2) communicate a descriptive norm of high participation, and (3) use the commitment-consistency principle by asking students to commit to participation. We find that none of the nudges were effective: all treatment effects are insignificant and close to zero in magnitude. Exploring heterogeneous treatment effects, we find evidence that the effectiveness of both the impact and commitment treatments differed across students. The impact treatment had a negative effect on the participation of bachelor-level students, but not on that of master-level students. The commitment treatment increased participation among students with good average grades, whereas it decreased participation for students whose average grades were poor.

Keywords: nudges; social norms; descriptive norm; commitment; student evaluation of teaching; participation; response rates; field experiment

JEL Classification: D90, D91, H41, I20

1. Introduction

Student evaluations of teaching (SETs) are widely used to measure teaching quality in higher education. The outcomes of such evaluations are consequential for both faculty and institutions: SETs affect faculty hiring, retention, promotion, and tenure decisions, as well as student enrollment numbers and the position of institutions in rankings and government audits (Becker and Watts 1999; Johnson 2000; Becker et al. 2012; Alter and Reback 2014).

A fundamental issue with SETs, which are increasingly being administered using online surveys, is that participation rates are often low (Dommeyer et al. 2004; Avery et al. 2006; Adams and Umbach 2012; for a review, see Spooren et al. 2013). These low response rates call into question the extent to which such evaluations reflect the overall opinion (Nulty 2008; Goos and Salomons 2017). Indeed, studies have found that respondents and non-respondents to SETs differ across a range of characteristics (Kherfi 2011; Adams and Umbach 2012; Spooren and Van Loon 2012; Goos and Salomons 2017). Several suggestions have been made to increase participation rates, of which explicit grade incentives have proven to be most effective (Johnson 2003; Dommeyer et al. 2004; Goodman et al. 2015; Sundstrom et al. 2016; Alvero et al. 2019; Lipsey and Shepperd 2020). However, many institutions and instructors are hesitant to adopt grade incentives out of ethical considerations or concerns about response validity. As a result, low participation rates remain a problem in many institutions.

In the present paper, we report on a large-scale randomized field experiment aimed at increasing participation rates in online SETs at the Erasmus School of Economics in the Netherlands. In our experiment, we nudged students to complete the online course evaluations by manipulating the email messages that invited them to do so. Influencing behavior through such "messaging nudges" has grown popular among policymakers, as these interventions can be easily implemented at virtually no additional cost using existing communication infrastructures.[1] In the context of SETs, messaging nudges may provide a feasible alternative to explicit grade incentives.

[1] Messaging interventions informed by behavioral insights have, for instance, been used by government agencies across the globe to affect outcomes in a broad set of policy areas. Examples include labor market and education decisions (invigorating job search by the unemployed, reducing attrition in adult education programs, increasing college enrollment of low-income students, reducing discrimination in hiring), economic growth (increasing uptake of government subsidies by small businesses), health and retirement choices (reducing smoking, increasing organ donations, expanding participation in retirement savings and health insurance plans, improving the efficiency of public hospitals, increasing early detection rates of cancer, and discouraging unnecessary prescription of antibiotics and controlled substances), and tax compliance, debt recovery, and collection of court

A large body of evidence demonstrates that simple, low-cost messaging nudges can increase prosocial behavior in a variety of domains, such as energy conservation (e.g., Nolan et al. 2008; Allcott 2011; Ayres et al. 2013; Allcott and Rogers 2014), environmental protection (e.g., Cialdini et al. 1990; Goldstein et al. 2008; Schultz et al. 2008), charitable giving (e.g., Shang and Croson 2009; Bartke et al. 2017), and tax compliance (e.g., Kettle et al. 2016; Hallsworth et al. 2017; Bott et al. 2020). Participation in SETs and, more generally, people's voluntary participation in surveys aimed at eliciting public opinion or satisfaction is another natural domain where messaging nudges can potentially be applied to increase prosocial behavior and thus improve the public good.[2]

In our experiment, we employed three popular messaging nudges to increase students' voluntary participation in course evaluations. In particular, the email messages that students received were designed to either (1) heighten the perceived impact of evaluations (impact treatment), (2) highlight that the majority of students complete the evaluations (peer treatment), or (3) use the commitment-consistency principle by asking students to commit to participation (commitment treatment). Students in the control treatment received an email containing routine information about when and where to provide their evaluations.[3]

Our paper makes contributions to three strands of literature. First, we add evidence from a field experiment to the literature on the effectiveness of messaging nudges in a to-date relatively unexplored application domain. As SETs and other evaluation surveys (e.g., customer satisfaction surveys) are increasingly being administered using online questionnaires that suffer from low participation rates, messaging nudges seem like a natural and promising way to increase participation. Second, our study contributes to the literature in higher education that aims to identify strategies for improving students' participation in teaching evaluations. To date, this literature has relied chiefly on observational or qualitative data (for a review, see, e.g., Nulty 2008). Existing experiments have focused on instructor-led interventions, particularly on the effect of offering grade incentives (e.g., Dommeyer et al. 2004; Sundstrom et al. 2016; Alvero et al. 2019; Lipsey and Shepperd 2020). Our study complements this literature by providing evidence for the effectiveness of messaging nudges sent by the

fines (The Behavioral Insights Team, 2016, 2017, 2019; National Science and Technology Council, Executive Office of the President, 2016).

[2] We are implicitly considering only unpaid participation in surveys, of which SETs are an example. Such participation can be argued to be a contribution to the public good, where the public good concerns the accuracy of the information elicited through surveys such as the SETs.

[3] This information was also present in the other three treatments (explained in Section 2). Hence, the treatments added particular pieces of information to the email that was used in the control treatment.

University administration to students directly. Such messages, which do not require the involvement of individual instructors, have the potential benefit that they can be more easily applied at scale than instructor-led interventions. Furthermore, the fact that the messaging nudges in our study have been applied independently of instructors rules out any possible selection or demand effects caused by instructors' support for the interventions or their awareness of the hypotheses being tested. Finally, our experimental setting allows us to study the heterogeneity of the effects of nudges along several factors shown to be significant correlates of SETs participation. We can, thus, contribute to the strand of literature that attempts to understand which types of nudging interventions work for whom and when (e.g., Allcott 2011; Chong et al. 2015; Costa and Kahn 2013).

Overall, we find that the messaging nudges did not increase students' participation in the course evaluations. Neither informing students about the impact of evaluations, nor indicating a descriptive norm of participation, nor inviting them to commit to participation positively impacted students' overall likelihood of completing their course evaluation surveys. Exploring heterogeneous treatment effects, we find evidence that the effectiveness of the impact and commitment treatments differed across students. The impact treatment decreased the participation of students in bachelor-level courses but not for students in master-level courses. The commitment treatment increased participation among students with good average grades, whereas it decreased participation for students whose grades were poor.

The paper proceeds as follows. Section 2 describes our experimental design and data. Our results are presented in section 3. Section 4 discusses our results and concludes. The appendix contains the details of our experimental treatments and supplementary analysis.

2. Experiment and Data

2.1 Experiment

The experiment took place in the 2013-2014 academic year at the Erasmus School of Economics (ESE) in collaboration with the school's Education Service Center.[4] Our

[4] For the type of social science experiment that we report on in this paper, ethical approval was not commonly required in the Netherlands at the time (in contrast to some other countries, such as the US). Therefore, we did not request ethical approval prior to conducting this experiment. It is worth emphasizing that the experiment did not fundamentally change the way SETs were conducted at the University. Instead, it only involved relatively small tweaks to the email messages inviting students to participate. The information provided to the students was always true; no deception took place.

experiment included all 3,485 bachelor and 1,552 master students enrolled at ESE and covered all 176 bachelor and 118 master courses offered during that academic year.

The academic year at ESE is organized into five blocks. Courses begin and end within a block. At the end of each block, students are asked to evaluate the quality of the courses that they attended by filling out an online survey. Participation in the evaluation surveys is voluntary, and participation rates are generally low, averaging roughly 25% for the ten blocks (two academic years) that preceded the year in which our experiment took place.

The evaluation surveys are administered through ESE's online platform for communication and course registration. One week before the end of a block, students receive an email message announcing the opening of the surveys and giving instructions about where to access the surveys. Students receive a second (identical) email message reminding them to fill out the surveys one week before closing. The surveys remain open for three weeks in total.

At the beginning of the academic year, all students registered on ESE's online platform were randomly assigned to one of the four treatments: control, impact, peer, and commitment.[5] Thus, the treatment assignment was on the student level. The treatment assignment remained unchanged throughout the year: students received the same type of messaging nudge throughout the five blocks of the academic year.

Our experimental treatments varied the text of the email messages sent to students, leaving the emails' timing and frequency unchanged. The email message in the control treatment contained a standard text informing students about the survey opening date, closing date, and where to find the surveys. The emails in the nudge treatments contained the same basic text information as in the control treatment but added particular pieces of information aimed at increasing participation rates. All treatment emails are reproduced in Appendix A (see Figures A1-A4).

The impact treatment was designed to stress the meaningfulness of the course evaluation. Students were informed that their feedback would be used to help improve teaching and to reward good lecturers. Meaning as an incentive has been found to lower reservation wage and increase labor supply in laboratory settings (Ariely et al. 2008), to improve job performance of charity fundraising callers (Grant 2008), and to increase job performance of students doing a data-entry job (Kosfeld et al. 2017). Research on student motivation for completing SETs

[5] This was done by first ranking the student numbers of all the registered students, and then assigning every four students into the four treatments respectively.

suggests that students' expectations concerning the impact of their evaluations (or lack thereof) are important for their decisions to participate (Spencer and Schmelkin 2002; Chen and Hoshower 2003). Although providing information about the impact of SETs has been frequently advised as a strategy for increasing participation rates (Johnson 2003; Nulty 2008; Chapman and Joines 2017), there has been little evidence of the extent to which such information increases participation rates. A notable exception is provided by Alvero et al. (2019), who, using a convenience sample, found that informing students about the importance of SETs did not increase participation rates. Our study tests this hypothesis using a large sample that includes all registered students at the Erasmus School of Economics.

Our peer treatment pointed out that in some courses over 80 percent of students participated in the SETs. Although participation rates in the evaluation surveys tend to be low overall, some courses have high participation rates. This allowed us to conduct the peer treatment without deception. The fraction of peers who evaluate a course is considered to be a descriptive norm. It has been well established that people tend to conform to descriptive norms, i.e., they follow the behavior that they perceive to be typical. Descriptive social norms have been found to influence cooperation and punishment in social dilemma games (Von Borgstede et al. 1999; Parks et al. 2001, Li et al. 2021), substance abuse among college students (for a review, see Perkins 2003), environmental and resource conservation (for a review, see Farrow et al. 2017), tax compliance (Kettle et al. 2016; Hallsworth et al. 2017), and voting (Gerber and Rogers 2009). In the context of survey participation, Misra et al. (2012) effectively used a descriptive social norm to increase participation in an evaluation survey among attendees of a scientific conference. Therefore, one would expect that students who learn that many of their fellow students participate in SETs would be more likely to participate themselves.

Our commitment treatment is based on the commitment-consistency principle, which states that people are more likely to behave in ways that are congruent with a position that they have previously endorsed (Aronson 1992; Cialdini 2007). In the commitment treatment, the first email served as the commitment device and asked students to indicate whether they intended to fill out the evaluation surveys. The second email was a simple reminder to perform the survey and was identical to the message sent in the control treatment. Commitment treatments similar to ours have been employed to stimulate healthy behaviors (e.g., Sandberg and Conner 2009; Bernstein et al. 2009) and voting (e.g., Greenwald et al. 1987; Smith et al. 2003; Mann 2005; Nickerson and Rogers 2010). These studies report either positive or insignificant treatment effects.

The commitment mechanism in our experiment technically allowed for a commitment not to participate (students could respond with a "no" to the question of whether they intended to participate). It was, therefore, in principle possible that the treatment would decrease participation. However, we hypothesized that the average participation rate should be higher in the commitment treatment than in the control treatment. Our hypothesis was grounded in the following logic: we expected that most students are not strongly opposed to participating in the course evaluations and that if students were to respond to the commitment message, they would overwhelmingly respond with a "yes" rather than a "no". Furthermore, we expected that students who had ignored the commitment request would not be less likely to participate than if they had received the standard message.

2.2 Data

Our primary outcome variable is whether a student completed the course evaluation survey. In particular, we look at every participation decision that every student had to make over the five blocks of the academic year in consideration. The number of times a particular student appears in our data set depends on the number of courses the student attended during the academic year. Because students sometimes do not attend a course they have registered for, we only consider courses in which they obtained a final grade (which could be a failing grade).[6]

In addition to whether or not each student completed the evaluation survey for each attended course, we gathered data on several student and course characteristics that are known to correlate with participation in SETs. In particular, we gathered data on students' grades for each course, whether the course was at the bachelor-level or the master-level, and the size of the course. Using the students' grades for each course, we computed their average grades in the academic year as a proxy for their academic ability. Furthermore, we computed the deviation of students' course grades from the students' own average grades for the academic year to investigate whether students are more or less likely to evaluate courses in which their performance is relatively good or bad. Prior research has shown that students' grade point average and course grade are positively related to participation in SETs (Layne et al. 1999;

[6] Students who did not obtain a final grade for a course were unlikely to evaluate the course. They only did so 10 percent of the time as compared to 23 percent for students who did obtain a final grade. If we conduct the analyses including all students who registered for the course, we obtain similar results (see online Appendix).

Table 1. Summary statistics

	Total			Mean (std. dev.) by treatment			
	Mean (std. dev.)	Min	Max	Control	Impact	Peer	Commit-ment
Average grade[(a)]	6.72 (1.18)	1.00	9.84	6.76 (1.20)	6.67 (1.20)	6.77 (1.15)	6.69 (1.18)
Course grade (deviation from average)[(b)]	0 (1.09)	-6.65	5.08	0 (1.06)	0 (1.14)	0 (1.07)	0 (1.09)
Course size	246 (192)	6	703	250 (192)	248 (193)	236 (186)	250 (195)
Master (1 if master-level; 0 if bachelor-level)	0.25 (0.43)	0	1	0.25 (0.43)	0.24 (0.43)	0.26 (0.44)	0.25 (0.44)
Number of observations[(c)]		30,221		7,775	7,432	7,520	7,494

Note:
[(a)] The assessment system in the Netherlands consists of grades from 1 (very poor) to 10 (outstanding). The grades 1 to 3, 9, and 10 are seldom given. A minimum grade of 5.5 is required to pass a course.
[(b)] Course grade is defined as the difference between the course grade and the student's average grade (course grade – average grade student). By construction, this variable sums to zero. Hence, the average per treatment (and the overall average) is zero.
[(c)] Because randomized treatment assignment was done at the student level and the number of courses varied across students, the number of observations is not equal across treatments.

Kherfi 2011; Spooren and Van Loon 2012; Reisenwitz 2016), and that participation in SETs is higher in more advanced courses (Spooren and Van Loon 2012). For course size, there is some evidence that participation rates are lower in larger courses (Goos and Salomons 2017).

Table 1 shows the summary statistics of our data. The average grade of a student deciding to complete a survey was 6.72 (out of 10). Course size varied between 6 and 703 students. Furthermore, 25 percent of the observations pertained to master-level courses, while 75 percent pertained to bachelor-level courses. Tests of equality of variable means across treatments provide no statistically significant evidence for imbalances in these observable characteristics (Average grade: clustered $F(3,4567) = 1.95,\ p = 0.119$; Course grade (deviation from average): clustered $F(3,4567) = 1.19, p = 0.311$; Course size: clustered $F(3,4567) = 1.63,$ $p = 0.180$; Master: clustered $Chi^2(3) = 1.601, p = 0.659$).

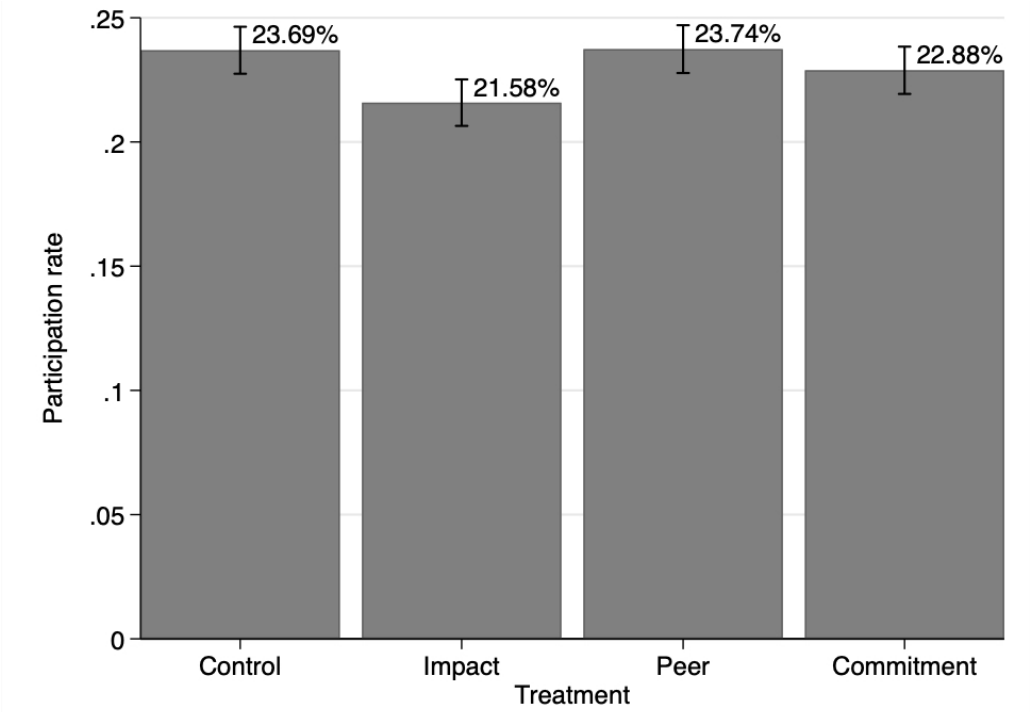
To further assess the degree to which the covariates are balanced between treatments, we consider normalized differences: the difference in averages between two treatments, scaled by the square root of the sum of the (within-treatment) variances. We compute the normalized differences between the control treatment and each of the three other treatments for all covariates. As a rule of thumb, Imbens and Rubin (2015) suggest that if important covariates have a normalized difference greater than 0.25, then simple regression methods may be unreliable. The largest absolute value of the normalized differences that we observe is below 0.1, suggesting that simple regression methods are sufficient to control for imbalances in covariates in our data. The following section will report results of logistic regression analyses that control for possible imbalances in the observable correlates of student participation in SETs.

3. Results

Figure 1 shows the students' rates of participation in the course evaluations across the four treatments. The participation rates in the four treatments were 23.69% (control), 21.58% (impact), 23.74% (peer), and 22.88% (commitment). In the commitment treatment, 12.21% of the students responded to the commitment email, of which 91.09% committed with a "yes". Participation rates did not differ significantly across the four treatments (clustered $Chi^2(3) = 3.016, p = 0.389$). In binary comparisons, none of the treatments differed significantly from the control treatment (impact vs. control: clustered $Chi^2(1) = 2.250, p = 0.134$; peer vs. control: clustered $Chi^2(1) = 0.001, p = 0.975$; commitment vs. control: clustered $Chi^2(1) = 0.332, p = 0.565$). Directionally, the participation rates in both the impact treatment and the commitment treatment were even a bit lower than that in the control treatment.

In order to control for (potential imbalances in) other factors that influence the participation rate and explore potential heterogeneous treatment effects, we conduct a logistic regression analysis. The dependent variable is a dummy variable that takes the value one if the student participated in the course evaluation and zero otherwise. We correct standard errors for clustering at the student level.

Figure 1. Participation rates by treatment



Note: The confidence bars represent the 95% confidence intervals.

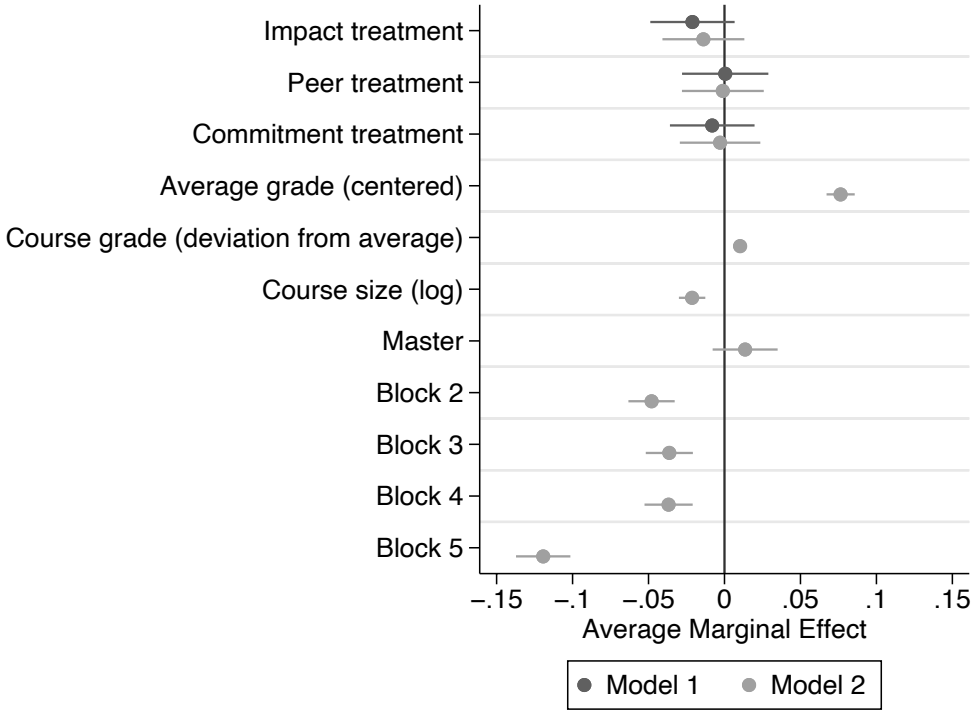
Table 2, Model 1 only includes the treatment dummies and confirms that none of our nudges significantly affected student participation in course evaluations (all $p > 0.135$). Table 2, Model 2 includes controls for the student's average grade (centered on the student-level average 6.54), the course grade (measured as the deviation of the student's grade for the course in question from the student's average grade), the course size (log-transformed), the course level (bachelor vs. master), and the block in which the course took place. Again, we find insignificant treatment effects for all nudges (all $p > 0.310$). In addition, we find that students with better average grades were more likely to complete course evaluations ($p < 0.001$). Moreover, the higher the course grade was relative to the student's average grade, the more likely it was that the student participated in the SET for that course ($p < 0.001$). Model 2 also shows that students were less likely to complete evaluations for larger courses ($p < 0.001$). Master-level students did not differ significantly from bachelor-level students in their likelihood to evaluate courses ($p = 0.209$). Participation rates were highest in the first block of the year and lowest in the fifth (last) block.

Table 2. Logit regression results

	Dependent variable: Completed (1 if the student completed course evaluation, 0 otherwise)		
	Model 1 coeff (p-value)	Model 2 coeff (p-value)	Model 3 coeff (p-value)
Impact	-0.120 (0.137)	-0.084 (0.313)	-0.183 (0.647)
Peer	0.003 (0.975)	-0.006 (0.937)	0.328 (0.410)
Commitment	-0.045 (0.571)	-0.017 (0.830)	0.236 (0.552)
Average grade (centered)		0.458 (0.000)	0.354 (0.000)
Course grade (deviation from average)		0.062 (0.000)	0.085 (0.000)
Course size (log)		-0.128 (0.000)	-0.081 (0.124)
Master (1 if master-level)		0.080 (0.209)	-0.096 (0.440)
Average grade * impact			0.063 (0.430)
Average grade * peer			0.132 (0.086)
Average grade * commitment			0.246 (0.002)
Course grade * impact			-0.032 (0.309)
Course grade * peer			-0.055 (0.070)
Course grade * commitment			-0.006 (0.846)
Course size * impact			-0.010 (0.894)
Course size * peer			-0.094 (0.204)
Course size * commitment			-0.078 (0.287)
Master * impact			0.409 (0.021)
Master * peer			0.244 (0.168)
Master * commitment			0.067 (0.707)
Block 2		-0.274 (0.000)	-0.274 (0.000)
Block 3		-0.204 (0.000)	-0.203 (0.000)
Block 4		-0.207 (0.000)	-0.204 (0.000)
Block 5		-0.773 (0.000)	-0.772 (0.000)
Constant	-1.170 (0.000)	-0.501 (0.002)	-0.639 (0.024)
Level of clustering	student	student	student
Number of observations	30,221	30,221	30,221
McFadden pseudo R-squared	0.0004	0.0535	0.0561

Because the logistic regression model coefficients are not easy to interpret, Figure 2 displays the average marginal treatment effects implied by Models 1 and 2. As shown in Figure 2, the marginal effects of our nudge treatments on students' likelihood of participation are close to zero across the board. By contrast, the marginal effects estimated for changes in the students' average grades, students' course grades, the course sizes, and the academic blocks are non-zero and tend to be larger in magnitude than the treatment effects.

Figure 2. Average marginal effects implied by Model 1 and Model 2

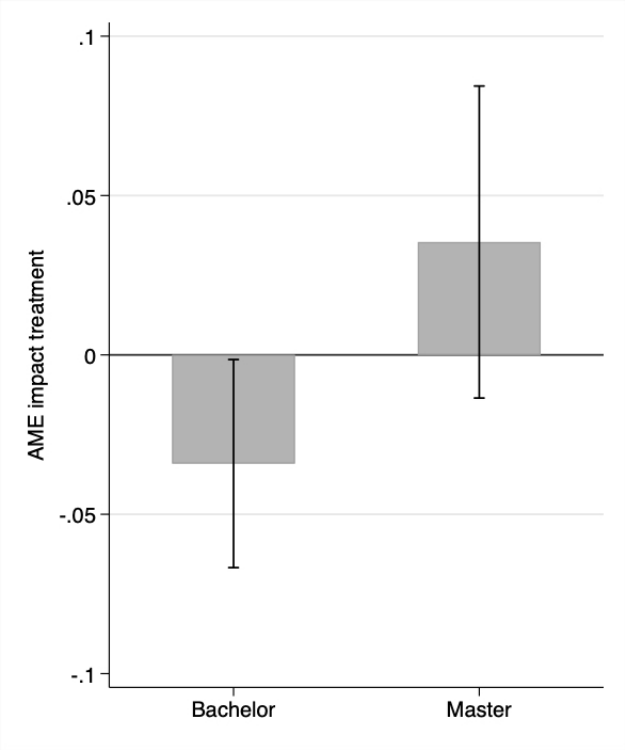


Note: Nodes show the average marginal effects on students' likelihood of participation in the course evaluations. Lines indicate the 95% confidence intervals for the marginal effects.

Table 2, Model 3 explores interactions of each of the nudge treatments with the average grade, the course grade, the course size, and the course level (master vs. bachelor). Overall, adding these twelve interactions significantly improves model performance ($Chi^2(12) = 24.86$, $p = 0.016$). Exploring interactions in more detail, we observe that course size did not significantly interact with any of the treatments (all $p > 0.200$). The course grade also did not significantly interact with the nudge messages (all $p > 0.070$). However, master-level students appear to have been more positively influenced by the impact treatment than students attending bachelor-level courses ($p = 0.021$). Finally, there is evidence that the commitment treatment was more effective for students with higher average grades ($p = 0.002$).

Given that we are simultaneously testing twelve different interactions, we need to account for multiple hypothesis testing. Both the interaction of the impact treatment with course level and the interaction of the commitment treatment with average grade remain significant after applying the Romano-Wolf multiple-hypothesis correction (Clarke et al. 2020; see also Romano et al. 2010). Figures 3 and 4 illustrate these interactions.

Figure 3. Impact treatment effects for bachelor-level and master-level courses (Model 3 estimates)

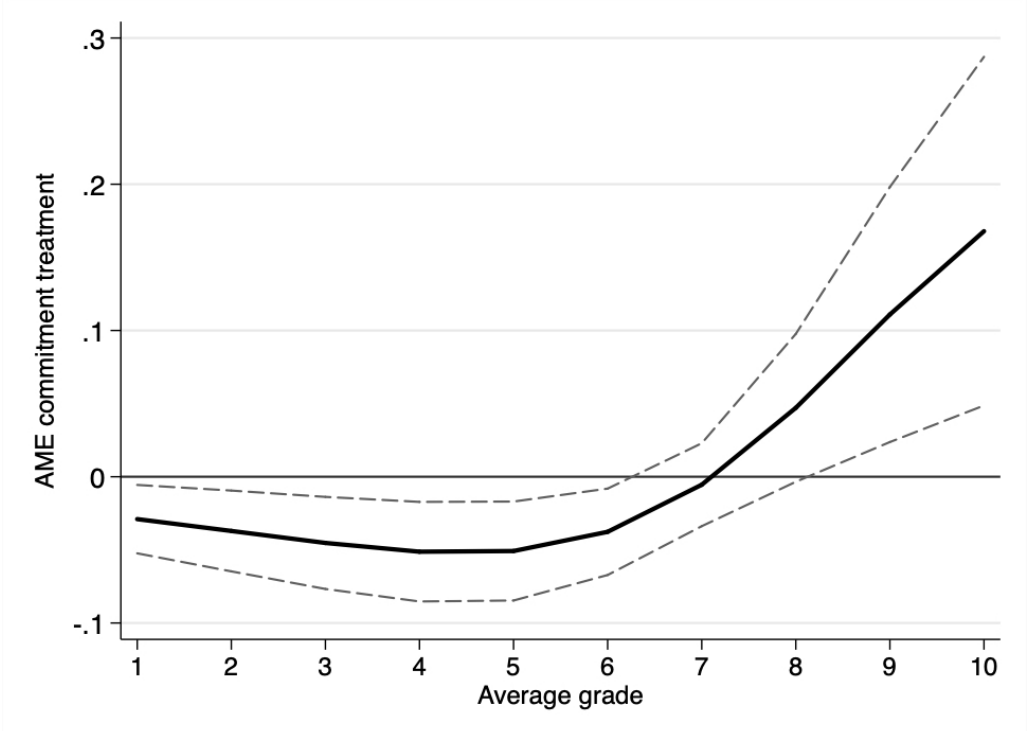


Note: The confidence bars represent the 95% confidence intervals. Estimates are obtained from Model 3 of Table 2.

Figure 3 plots the estimated average effects of the impact treatment for bachelor vs. master-level courses. We observe that the estimated effect size is negative for students attending bachelor-level courses but positive for students attending master-level courses. However, only the negative effect for bachelor students is statistically significant (exact estimates are in Appendix B, Table B1).

Figure 4 plots the estimated average effects of the commitment treatment as a function of the student's average grade. We observe that the commitment treatment increased the predicted likelihood of participation of students with good average grades. In contrast, the commitment treatment decreased the participation likelihood of students with poor average grades (exact estimates are in Appendix B, Table B2). Exploring the commitment treatment in more detail, we observe that students with higher average grades responded more positively at each stage of the commitment treatment process. Compared to students with poor average grades, students with high grades were: (i) more likely to respond to the commitment email; (ii) conditionally on responding, more likely to respond in the affirmative; and (3) more likely to meet their commitment and complete the course evaluation after they had committed (see Appendix B, Table B3).

Figure 4. Commitment treatment effects as a function of average grade (Model 3 estimates)



Note: Solid line plots the estimated marginal effects of the commitment treatment dummy. Dashed lines indicate the 95% confidence intervals. Estimates are obtained from Model 3 of Table 2.

4. Discussion and Conclusion

We conducted a field experiment that tests the efficacy of three low-cost messaging nudges in encouraging students' participation in the online evaluation of their courses. Our nudges were designed to enhance students' perceived impact of their participation, to signal that participation was the descriptive norm, and to elicit a commitment to participate from students. Overall, we find that none of the nudges significantly increased participation rates.

Our finding that impact information was insignificant in raising participation in course evaluations is consistent with that of Alvero et al. (2019)—the only other experimental study that we are aware of, which tested the efficacy of impact information in raising student participation in SETs. An exploration of heterogeneous treatment effects further suggested that the impact information may have even had a negative effect on the participation of students attending bachelor-level courses, but not on that of students attending master-level courses. These results are surprising given that students who are asked about their motivation to complete SETs often state that the perceived impact is an important motivator (Spencer and Schmelkin 2002; Chen and Hoshower 2003). One interpretation is that providing students with

simple impact information is insufficient to alter students' perceived impact of SETs. Alternatively, it may also be that students are not actually as motivated by the impact of their course evaluations as they say they are, echoing a similar dissonance between claimed and revealed importance of the factors that motivate pro-environmental behavior (Nolan et al. 2008).

To the best of our knowledge, this paper is the first to explore the power of descriptive norms to raise participation in SETs. We do not find descriptive norms to be effective in raising student participation in evaluations. Our study is not the first to find null effects for nudges relying on descriptive norms. For instance, Chabé-Ferret et al. (2019) used descriptive norms to nudge farmers in south-western France to reduce their irrigation water consumption and found the overall treatment effects to be insignificant (for reviews, see Farrow et al. 2017; Hummel and Maedche 2019).

The consistency principle underlying our commitment treatment is well-known in psychology (Festinger 1957; Aronson 1992). The commitment treatment has been shown to influence behavior in various domains (for a review, see Wilding et al. 2016). However, to date, the principle has not been tested for its efficacy in increasing survey participation. Our overall finding is that the commitment treatment did not significantly increase students' participation in the course evaluation surveys. Exploring heterogeneous treatment effects, we find evidence that the effectiveness of the commitment treatment differed across students: it increased participation of students with good average grades, whereas it decreased participation for students whose grades were poor.

In general, a commitment treatment's efficacy relies on sufficiently high commitment rates, in addition to the strength of the commitment-consistency principle. Previous studies reporting significant effects of the commitment treatment often show high rates of commitment. For instance, Baca-Motes et al. (2013) conducted a field experiment showing a significant commitment effect on hotel guests' pro-environmental behavior. In their experiment, commitment rates were high: over 80 percent of all targeted guests agreed to commit. In our experiment, the commitment rate among students was 11 percent (the response rate to the commitment email was 12 percent, and 91 percent of the responders committed with a "yes"). This low commitment rate was likely an important reason for our overall null effect. It is well possible that alternative implementations of the commitment treatment, in particular an implementation in which students commit to participation in the SETs in a personal interaction

with the instructor, would produce higher commitment rates. However, such an implementation would also be significantly more costly, as it requires a time investment from the instructor to obtain commitments from students personally. Our results highlight a possible drawback of commitment nudges: they rely heavily on the initial commitment, and such a commitment may be harder to obtain than previously discussed.

An issue that underlies all messaging-based nudge treatments is the difficulty of ensuring that the target audience reads the message. Only those who read the message can be affected by it, and the observed treatment effect will therefore be proportional to the rate by which the audience reads the message. In our study, the messaging nudges were incorporated into formal email messages from the university. The university uses such emails to communicate all essential administrative information to students. Therefore, widespread inattention to such emails seems unlikely. However, we cannot rule out that a non-negligible fraction of students ignore these formal emails (potentially after a quick screening to judge the importance of the email) and that inattention to the nudge messages is one of the mechanisms that is driving the overall null results that we observe.

It is worth emphasizing that our experiment was well powered to detect even small (observed) effect sizes. For instance, if the true overall effect size was just 2.5 percentage points (corresponding to a situation where 50% of students ignore the nudge message and where the nudge effect size is 5 percentage points), we would have a 95 percent chance of rejecting the null hypothesis at the five percent level. If the true overall effect size was only 2 percentage points, we still would have a probability of over 80 percent of rejecting the null at the five percent level. Thus, the fact that we find null effects on aggregate suggests that the messaging nudges were not effective in this setting.

Our results show that nudges that have been documented to work well in the literature were not effective when applied to increase student participation in SETs. In doing so, our study adds to evidence suggesting that nudges may not always be as effective as sometimes claimed in the literature. Recently, DellaVigna and Linos (2020) reviewed evidence from all published and unpublished large-scale nudge trials conducted by two major nudge units in the United States. Comparing the nudge effects found in these large-scale trials to the effects of the nudges documented in the academic literature, the authors find that the average effect sizes in the large-scale field trials are much smaller than those reported in the literature and that publication bias explains a large share of the gap. Sanders et al. (2018) describe the systemic barriers that

discourage the publication of field studies conducted in government agencies, especially when these concern null results.

The efficacies of particular nudge interventions likely depend on a host of factors, including the characteristics of the group being nudged and the behavioral domain being targeted. Ultimately, a deeper understanding of which types of nudges work under which conditions can only be achieved when sufficient evidence from more varied groups and behavioral domains accumulates. Our results add to this evidence.

References

Adams, Meredith JD, and Paul D. Umbach. "Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments." *Research in Higher Education* 53, no. 5 (2012): 576-591. DOI: 10.1007/s11162-011-9240-5

Allcott, Hunt. "Social norms and energy conservation." *Journal of Public Economics* 95, no. 9-10 (2011): 1082-1095. DOI: 10.1016/j.jpubeco.2011.03.003

Allcott, Hunt, and Todd Rogers. "The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation." *American Economic Review* 104, no. 10 (2014): 3003-37. DOI: 10.1257/aer.104.10.3003

Alter, Molly, and Randall Reback. "True for your school? How changing reputations alter demand for selective US colleges." *Educational Evaluation and Policy Analysis* 36, no. 3 (2014): 346-370. DOI: 10.3102/0162373713517934

Alvero, Alicia M., Kathleen Mangiapanello, and Jennifer Valad. "The effects of incentives, instructor motivation and feedback strategies on faculty evaluation response rates in large and small class sizes." *Assessment and Evaluation in Higher Education* 44, no. 4 (2019): 501-515. DOI: 10.1080/02602938.2018.1521913

Ariely, Dan, Emir Kamenica, and Dražen Prelec. "Man's search for meaning: The case of Legos." *Journal of Economic Behavior and Organization* 67, no. 3-4 (2008): 671-677. DOI: 10.1016/j.jebo.2008.01.004

Aronson, Elliot. "The return of the repressed: Dissonance theory makes a comeback." *Psychological Inquiry* 3, no. 4 (1992): 303-311. DOI: 10.1207/s15327965pli0304_1

Avery, Rosemary J., W. Keith Bryant, Alan Mathios, Hyojin Kang, and Duncan Bell. "Electronic course evaluations: does an online delivery system influence student evaluations?" *Journal of Economic Education* 37, no. 1 (2006): 21-37. DOI: 10.3200/JECE.37.1.21-37

Ayres, Ian, Sophie Raseman, and Alice Shih. "Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage." *Journal of Law, Economics, and Organization* 29, no. 5 (2013): 992-1022. DOI: 10.1093/jleo/ews020

Baca-Motes, Katie, Amber Brown, Ayelet Gneezy, Elizabeth A. Keenan, and Leif D. Nelson. "Commitment and behavior change: Evidence from the field." *Journal of Consumer Research* 39, no. 5 (2013): 1070-1084. DOI: 10.1086/667226

Bartke, Simon, Andreas Friedl, Felix Gelhaar, and Laura Reh. "Social comparison nudges—Guessing the norm increases charitable giving." *Economics Letters* 152 (2017): 73-75. DOI: 10.1016/j.econlet.2016.12.023

Becker, William E., William Bosshardt, and Michael Watts. "How departments of economics evaluate teaching." *Journal of Economic Education* 43, no. 3 (2012): 325-333. DOI: 10.1080/00220485.2012.686826

Becker, William E., and Michael Watts. "How departments of economics evaluate teaching." *American Economic Review* 89, no. 2 (1999): 344-349. DOI: 10.1257/aer.89.2.344

Bernstein, Edward, Erika Edwards, David Dorfman, Tim Heeren, Caleb Bliss, and Judith Bernstein. "Screening and brief intervention to reduce marijuana use among youth and young adults in a pediatric emergency department." *Academic Emergency Medicine* 16, no. 11 (2009): 1174-1185. DOI: 10.1111/j.1553-2712.2009.00490.x

Bott, Kristina M., Alexander W. Cappelen, Erik Ø. Sørensen, and Bertil Tungodden. "You've got mail: A randomized field experiment on tax evasion." *Management Science* 66, no.7 (2020). DOI: 10.1287/mnsc.2019.3390

Chabé-Ferret, Sylvain, Philippe Le Coent, Arnaud Reynaud, Julie Subervie, and Daniel Lepercq. "Can we nudge farmers into saving water? Evidence from a randomised experiment." *European Review of Agricultural Economics* 46, no. 3 (2019): 393-416. DOI: 10.1093/erae/jbz022

Chapman, Diane D., and Jeffrey A. Joines. “Strategies for increasing response rates for online end-of-course evaluations.” *International Journal of Teaching and Learning in Higher Education* 29, no. 1 (2017): 47-60. Available online.

Chen, Yining, and Leon B. Hoshower. “Student evaluation of teaching effectiveness: An assessment of student perception and motivation.” *Assessment and Evaluation in Higher Education* 28, no. 1 (2003): 71-88. DOI: 10.1080/02602930301683

Chong, Alberto, Dean Karlan, Jeremy Shapiro, and Jonathan Zinman. “(Ineffective) messages to encourage recycling: evidence from a randomized evaluation in Peru.” *The World Bank Economic Review* 29, no. 1 (2015): 180-206. DOI: 10.1093/wber/lht022

Cialdini, Robert B. *Influence: The psychology of persuasion*. New York: Collins, 2007.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren. “A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places.” *Journal of Personality and Social Psychology* 58, no. 6 (1990): 1015. DOI: 10.1037/0022-3514.58.6.1015

Clarke, Damian, Joseph P. Romano, and Michael Wolf. “The Romano–Wolf multiple-hypothesis correction in Stata.” *The Stata Journal* 20, no. 4 (2020): 812-843. DOI: 10.1177/1536867X20976314

Costa, Dora L., and Matthew E. Kahn. “Energy conservation ‘nudges’ and environmentalist ideology: Evidence from a randomized residential electricity field experiment.” *Journal of the European Economic Association* 11, no. 3 (2013): 680-702. DOI: 10.1111/jeea.12011

DellaVigna, Stefano, and Elizabeth Linos. “RCTs to scale: Comprehensive evidence from two nudge units.” *NBER Working Paper* 27594 (2020). Available online.

Dommeyer, Curt J., Paul Baum, Robert W. Hanna, and Kenneth S. Chapman. “Gathering Faculty Teaching Evaluations by in-class and Online Surveys: Their Effects on Response Rates and Evaluations.” *Assessment and Evaluation in Higher Education* 29, no. 5 (2004): 611-623. DOI: 10.1080/02602930410001689171

Farrow, Katherine, Gilles Grolleau, and Lisette Ibanez. “Social norms and pro-environmental behavior: A review of the evidence.” *Ecological Economics* 140 (2017): 1-13. DOI: 10.1016/j.ecolecon.2017.04.017

Festinger, Leon. *A theory of cognitive dissonance*. Vol. 2. Stanford university press, 1957.

Gerber, Alan S., and Todd Rogers. "Descriptive social norms and motivation to vote: Everybody's voting and so should you." *Journal of Politics* 71, no. 1 (2009): 178-191. DOI: 10.1017/S0022381608090117

Goldstein, Noah J., Robert B. Cialdini, and Vladas Griskevicius. "A room with a viewpoint: Using social norms to motivate environmental conservation in hotels." *Journal of Consumer Research* 35, no. 3 (2008): 472-482. DOI: 10.1086/586910

Goodman, James, Robert Anson, and Marcia Belcheir. "The effect of incentives and other instructor-driven strategies to increase online student evaluation response rates." *Assessment and Evaluation in Higher Education* 40, no. 7 (2015): 958-970. DOI: 10.1080/02602938.2014.960364

Goos, Maarten, and Anna Salomons. "Measuring teaching quality in higher education: assessing selection bias in course evaluations." *Research in Higher Education* 58, no. 4 (2017): 341-364. DOI: 10.1007/s11162-016-9429-8

Grant, Adam M. "Employees without a cause: The motivational effects of prosocial impact in public service." *International Public Management Journal* 11, no. 1 (2008): 48-66. DOI: 10.1080/10967490801887905

Greenwald, Anthony G., Catherine G. Carnot, Rebecca Beach, and Barbara Young. "Increasing voting behavior by asking people if they expect to vote." *Journal of Applied Psychology* 72, no. 2 (1987): 315. DOI: 10.1037/0021-9010.72.2.315

Hallsworth, Michael, John A. List, Robert D. Metcalfe, and Ivo Vlaev. "The behavioralist as tax collector: Using natural field experiments to enhance tax compliance." *Journal of Public Economics* 148 (2017): 14-31. DOI: 10.1016/j.jpubeco.2017.02.003

Hummel, Dennis, and Alexander Maedche. "How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies." *Journal of Behavioral and Experimental Economics* 80 (2019): 47-58. DOI: 10.1016/j.socec.2019.03.005

Imbens, Guido W., and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press (2015).

Johnson, Rachel. “The authority of the student evaluation questionnaire.” *Teaching in Higher Education* 5, no. 4 (2000): 419-434. DOI: 10.1080/713699176

Johnson, Trav D. “Online Student Ratings: Will Students Respond?” *New Directions for Teaching and Learning* 96 (2003): 49-59. DOI: 10.1002/tl.122

Kettle, Stewart, Marco Hernandez, Simon Ruda, and Michael Sanders. *Behavioral interventions in tax compliance: Evidence from Guatemala*. The World Bank (2016). Available online.

Kherfi, Samer. “Whose opinion is it anyway? Determinants of participation in student evaluation of teaching.” *Journal of Economic Education* 42, no. 1 (2011): 19-30. DOI: 10.1080/00220485.2011.536487

Kosfeld, Michael, Susanne Neckermann, and Xiaolan Yang. “The effects of financial and recognition incentives across work contexts: The role of meaning.” *Economic Inquiry* 55, no. 1 (2017): 237-247. DOI: 10.1111/ecin.12350

Layne, Benjamin H., Joseph R. DeCristoforo, and Dixie McGinty. “Electronic versus traditional student ratings of instruction.” *Research in Higher Education* 40, no. 2 (1999): 221-232. DOI: 10.1023/A:1018738731032

Li, Xueheng, Lucas Molleman, and Dennie van Dolder. “Do descriptive social norms drive peer punishment? Conditional punishment strategies and their impact on cooperation.” *Evolution and Human Behavior* 42, no. 5 (2021): 469-479. DOI: 10.1016/j.evolhumbehav.2021.04.002

Lipsey, Nikolette, and James Shepperd. “Examining strategies to increase student evaluation of teaching completion rates.” *Assessment and Evaluation in Higher Education* (2020). DOI: 10.1080/02602938.2020.1782343

Mann, Christopher B. “Unintentional voter mobilization: Does participation in preelection surveys increase voter turnout?.” *Annals of the American Academy of Political and Social Science* 601, no. 1 (2005): 155-168. DOI: 10.1177/0002716205278151

Misra, Shalini, Daniel Stokols, and Anne Heberger Marino. “Using norm–based appeals to increase response rates in evaluation research: A field experiment.” *American Journal of Evaluation* 33, no. 1 (2012): 88-98. DOI: 10.1177/1098214011414862

National Science and Technology Council, Executive Office of the President. *Social and Behavioral Science Team 2016 Annual Report*. 2016. Available online.

Nickerson, David W., and Todd Rogers. "Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making." *Psychological Science* 21, no. 2 (2010): 194-199. DOI: 10.1177/0956797609359326

Nolan, Jessica M., P. Wesley Schultz, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. "Normative social influence is underdetected." *Personality and Social Psychology Bulletin* 34, no. 7 (2008): 913-923. DOI: 10.1177/0146167208316691

Nulty, Duncan D. "The adequacy of response rates to online and paper surveys: what can be done?" *Assessment and Evaluation in Higher Education* 33, no. 3 (2008): 301-314. DOI: 10.1080/02602930701293231

Parks, Craig D., Lawrence J. Sanna, and Susan R. Berel. "Actions of similar others as inducements to cooperate in social dilemmas." *Personality and Social Psychology Bulletin* 27, no. 3 (2001): 345-354. DOI: 10.1177/0146167201273008

Perkins, H. Wesley. "The emergence and evolution of the social norms approach to substance abuse prevention." *The social norms approach to preventing school and college age substance abuse: A handbook for educators, counselors, and clinicians* (2003): 3-17.

Reisenwitz, Timothy H. "Student evaluation of teaching: An investigation of nonresponse bias in an online context." *Journal of Marketing Education* 38, no. 1 (2016): 7-17. DOI: 10.1177/0273475315596778

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. "Hypothesis testing in econometrics." *Annual Review of Economics* 2, no. 1 (2010): 75-104. DOI: 10.1146/annurev.economics.102308.124342

Sandberg, Tracy, and Mark Conner. "A mere measurement effect for anticipated regret: Impacts on cervical screening attendance." *British Journal of Social Psychology* 48, no. 2 (2009): 221-236. DOI: 10.1348/014466608X347001

Sanders, Michael, Veerle Snijders, and Michael Hallsworth. "Behavioural science and policy: where are we now and where are we going?." *Behavioural Public Policy* 2, no. 2 (2018): 144-167. DOI: 10.1017/bpp.2018.17

Schultz, Wesley P., Azar M. Khazian, and Adam C. Zaleski. “Using normative social influence to promote conservation among hotel guests.” *Social Influence* 3, no. 1 (2008): 4-23. DOI: 10.1080/15534510701755614

Shang, Jen, and Rachel Croson. “A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods.” *The Economic Journal* 119, no. 540 (2009): 1422-1439. DOI: 10.1111/j.1468-0297.2009.02267.x

Smith, Jennifer K., Alan S. Gerber, and Anton Orlich. “Self-prophecy effects and voter turnout: An experimental replication.” *Political Psychology* 24, no. 3 (2003): 593-604. DOI: 10.1111/0162-895X.00342

Spencer, Karin J., and Liora Pedhazur Schmelkin. “Student perspectives on teaching and its evaluation.” *Assessment and Evaluation in Higher Education* 27, no. 5 (2002): 397-409. DOI: 10.1080/0260293022000009285

Spooren, Pieter, Bert Brockx, and Dimitri Mortelmans. “On the validity of student evaluation of teaching: The state of the art.” *Review of Educational Research* 83, no. 4 (2013): 598-642. DOI: 10.3102/0034654313496870

Spooren, Pieter, and Francis Van Loon. “Who participates (not)? A non-response analysis on students’ evaluations of teaching.” *Procedia – Social and Behavioral Sciences* 69 (2012): 990-996. DOI: 10.1016/j.sbspro.2012.12.025

Sundstrom, Eric D., Erin E. Hardin, and Matthew J. Shaffer. “Extra credit micro-incentives and response rates for online course evaluations: Two quasi-experiments.” *Teaching of Psychology* 43, no. 4 (2016): 276-284. DOI: 10.1177/0098628316662754

The Behavioural Insights Team. *Update Report 2015–16*. 2016. Available online.

The Behavioural Insights Team. *Update Report 2016–17*. 2017. Available online.

The Behavioural Insights Team. *Annual Report 2017–18*. 2019. Available online.

Von Borgstede, Chris, Ulf Dahlstrand, and Anders Biel. “From Ought to Is: Moral Norms in Large-Scale Social Dilemmas.” *Goteborg Psychological Reports* 29, no. 5 (1999): 1-17. Available online.

Wilding, Sarah, Mark Conner, Tracy Sandberg, Andrew Prestwich, Rebecca Lawton, Chantelle Wood, Eleanor Miles, Gaston Godin, and Paschal Sheeran. “The question-behaviour effect: a theoretical and methodological review and meta-analysis.” *European Review of Social Psychology* 27, no. 1 (2016): 196-230. DOI: 10.1080/10463283.2016.1245940

Appendix A. Treatment Email Messages

Figure A1. Email message in the control treatment

Note: Students received two emails identical in content. The first email was sent at the opening of the evaluation surveys. The second (reminder) email was sent one week before the closing of the surveys.

FOR ENGLISH SCROLL DOWN

Beste student,

Je kunt de kwaliteit van jouw vakken evalueren door met je ERNA gebruikersnaam en wachtwoord in te loggen op SIN-Online. Klik op *My Page*, scroll omlaag en klik op *Questionnaires Waiting*.

Het invullen van de vragenlijst zal slechts 5 minuten in beslag nemen. De vragenlijst is beschikbaar tot een week na het tentamen.

Bij voorbaat dank voor je deelname,

Ria Koolen, Onderwijs Service Centrum, ESE

Dear student,

You can evaluate the quality of your courses by logging into SIN-Online using your ERNA username and password. Click on *My Page*, scroll down, then click on *Questionnaires Waiting*.

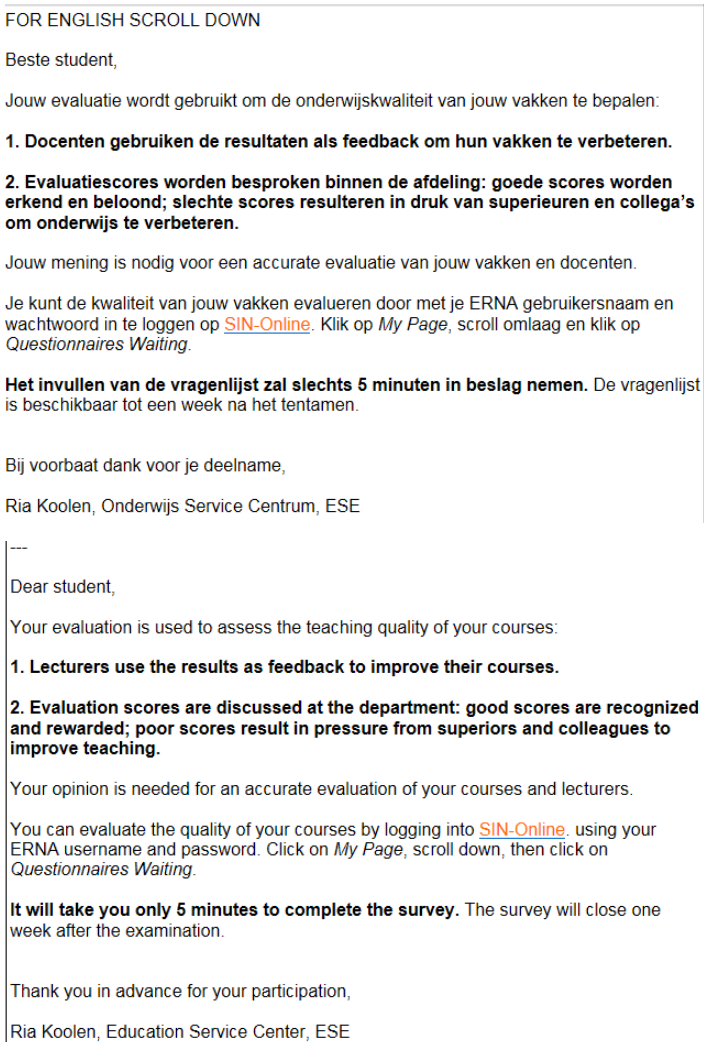
It will take you only 5 minutes to complete the survey. The survey will close one week after the examination.

Thank you in advance for your participation,

Ria Koolen, Education Service Center, ESE

Figure A2. Email message in the impact treatment

Note: Students received two emails identical in content. The first email was sent at the opening of the evaluation surveys. The second (reminder) email was sent one week before the closing of the surveys.



FOR ENGLISH SCROLL DOWN

Beste student,

Jouw evaluatie wordt gebruikt om de onderwijskwaliteit van jouw vakken te bepalen:

1. Docenten gebruiken de resultaten als feedback om hun vakken te verbeteren.

2. Evaluatiescores worden besproken binnen de afdeling: goede scores worden erkend en beloond; slechte scores resulteren in druk van superieuren en collega's om onderwijs te verbeteren.

Jouw mening is nodig voor een accurate evaluatie van jouw vakken en docenten.

Je kunt de kwaliteit van jouw vakken evalueren door met je ERNA gebruikersnaam en wachtwoord in te loggen op SIN-Online. Klik op *My Page*, scroll omlaag en klik op *Questionnaires Waiting*.

Het invullen van de vragenlijst zal slechts 5 minuten in beslag nemen. De vragenlijst is beschikbaar tot een week na het tentamen.

Bij voorbaat dank voor je deelname,

Ria Koolen, Onderwijs Service Centrum, ESE

Dear student,

Your evaluation is used to assess the teaching quality of your courses:

1. Lecturers use the results as feedback to improve their courses.

2. Evaluation scores are discussed at the department: good scores are recognized and rewarded; poor scores result in pressure from superiors and colleagues to improve teaching.

Your opinion is needed for an accurate evaluation of your courses and lecturers.

You can evaluate the quality of your courses by logging into SIN-Online, using your ERNA username and password. Click on *My Page*, scroll down, then click on *Questionnaires Waiting*.

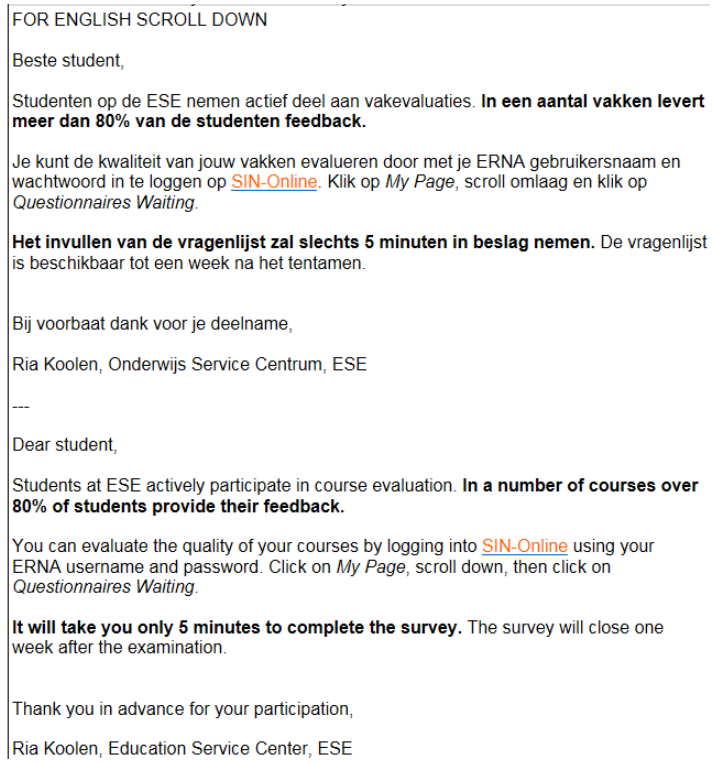
It will take you only 5 minutes to complete the survey. The survey will close one week after the examination.

Thank you in advance for your participation,

Ria Koolen, Education Service Center, ESE

Figure A3. Email message in the peer treatment

Note: Students received two emails identical in content. The first email was sent at the opening of the evaluation surveys. The second (reminder) email was sent one week before the closing of the surveys.



FOR ENGLISH SCROLL DOWN

Beste student,

Studenten op de ESE nemen actief deel aan vakevaluaties. **In een aantal vakken levert meer dan 80% van de studenten feedback.**

Je kunt de kwaliteit van jouw vakken evalueren door met je ERNA gebruikersnaam en wachtwoord in te loggen op SIN-Online. Klik op *My Page*, scroll omlaag en klik op *Questionnaires Waiting*.

Het invullen van de vragenlijst zal slechts 5 minuten in beslag nemen. De vragenlijst is beschikbaar tot een week na het tentamen.

Bij voorbaat dank voor je deelname,

Ria Koolen, Onderwijs Service Centrum, ESE

Dear student,

Students at ESE actively participate in course evaluation. **In a number of courses over 80% of students provide their feedback.**

You can evaluate the quality of your courses by logging into SIN-Online using your ERNA username and password. Click on *My Page*, scroll down, then click on *Questionnaires Waiting*.

It will take you only 5 minutes to complete the survey. The survey will close one week after the examination.

Thank you in advance for your participation,

Ria Koolen, Education Service Center, ESE

Figure A4. Email message in the commitment treatment

Note: Students received two emails. The first email (A) was sent at the opening of the evaluation surveys. The second (reminder) email (B) was sent one week before the closing of the surveys.

(A) First email in commitment treatment

FOR ENGLISH SCROLL DOWN

Beste student,

Het is voor ons nuttig om vooraf te weten hoe veel evaluatievragenlijsten zullen worden ingevuld. **Laat ons alsjeblieft weten of je van plan bent om de vakevaluatie in te vullen, door "yes" of "no" te klikken na de volgende link: Link.**

Je kunt de kwaliteit van jouw vakken evalueren door met je ERNA gebruikersnaam en wachtwoord in te loggen op SIN-Online. Klik op *My Page*, scroll omlaag en klik op *Questionnaires Waiting*.

Het invullen van de vragenlijst zal slechts 5 minuten in beslag nemen. De vragenlijst is beschikbaar tot een week na het tentamen.

Bij voorbaat dank voor je deelname,

Ria Koolen, Onderwijs Service Centrum, ESE

Dear student,

It is helpful for us to know in advance how many evaluation surveys will be completed. **Please tell us whether you intend to complete the course evaluation, by clicking "yes" or "no" using the following link: Link.**

You can evaluate the quality of your courses by logging into SIN-Online using your ERNA username and password. Click on *My Page*, scroll down, then click on *Questionnaires Waiting*.

It will take you only 5 minutes to complete the survey. The survey will close one week after the examination.

Thank you in advance for your participation,

Ria Koolen, Education Service Center, ESE

(B) Second email in commitment treatment (identical to email in control treatment)

FOR ENGLISH SCROLL DOWN

Beste student,

Je kunt de kwaliteit van jouw vakken evalueren door met je ERNA gebruikersnaam en wachtwoord in te loggen op SIN-Online. Klik op *My Page*, scroll omlaag en klik op *Questionnaires Waiting*.

Het invullen van de vragenlijst zal slechts 5 minuten in beslag nemen. De vragenlijst is beschikbaar tot een week na het tentamen.

Bij voorbaat dank voor je deelname,

Ria Koolen, Onderwijs Service Centrum, ESE

Dear student,

You can evaluate the quality of your courses by logging into SIN-Online using your ERNA username and password. Click on *My Page*, scroll down, then click on *Questionnaires Waiting*.

It will take you only 5 minutes to complete the survey. The survey will close one week after the examination.

Thank you in advance for your participation,

Ria Koolen, Education Service Center, ESE

Appendix B. Supplementary Analyses

Table B1. Impact treatment effects (on the likelihood of participation) as a function of course level

		Avg. marginal effect in percentage points (p-value)
Master	0	-3.410 (0.041)
	1	3.541 (0.156)

Table B2. Commitment treatment effects (on the likelihood of participation) as a function of average grade

		Avg. marginal effect in percentage points (p-value)
Average grade	1	-2.897 (0.015)
	2	-3.706 (0.009)
	3	-4.525 (0.005)
	4	-5.124 (0.003)
	5	-5.078 (0.003)
	6	-3.764 (0.012)
	7	-0.546 (0.706)
	8	4.711 (0.068)
	9	11.094 (0.013)
	10	16.790 (0.006)

Table B3. Logit regression results for models exploring the mechanisms within the commitment treatment

	Dependent variable			
coeff (p-value)	Responded to commitment email	Responded with "yes" (if responded)	Completed evaluation (if responded with "yes")	Completed evaluation (if did not respond to commitment email)
Average grade (centered)	0.581 (0.000)	0.261 (0.072)	0.423 (0.001)	0.490 (0.000)
Course grade (deviation from average grade)	-0.007 (0.750)	0.092 (0.409)	-0.067 (0.426)	0.128 (0.000)
Course size (log)	-0.004 (0.957)	-0.381 (0.090)	-0.144 (0.258)	-0.208 (0.000)
Master (1 if master-level)	-0.332 (0.062)	0.289 (0.573)	0.079 (0.789)	0.187 (0.169)
Block 2	-0.432 (0.000)	0.291 (0.431)	0.273 (0.305)	-0.410 (0.003)
Block 3	-0.775 (0.000)	0.625 (0.201)	1.379 (0.000)	0.089 (0.460)
Block 4	-0.849 (0.000)	0.171 (0.743)	1.450 (0.001)	0.039 (0.748)
Block 5	-1.265 (0.000)	0.426 (0.628)	1.022 (0.134)	-0.410 (0.020)
Constant	-1.463 (0.002)	3.962 (0.003)	1.498 (0.034)	-0.858 (0.006)
Level of clustering	student	student	student	student
Number of observations	7,494	1,014	927	6,480
McFadden pseudo R-squared	0.0703	0.0475	0.0908	0.0643

Online Appendix of
"Nudging Student Participation in Online Evaluations of Teaching: Evidence from a Field Experiment"

Susanne Neckermann[a], Uyanga Turmunkh (corresponding author)[b,c], Dennie van Dolder[d,e], Tong V. Wang[f]

[a] Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, USA, sneckermann@uchicago.edu; [b] Department of Economics and Quantitative Methods, IESEG School of Management, 3 Rue de la Digue, 59000 Lille, France, u.turmunkh@ieseg.fr; [c] LEM-CNRS 9221; [d] School of Business and Economics, Vrije Universiteit (VU) Amsterdam, Gustav Mahlerplein 117, 1082 MS Amsterdam, the Netherlands, d.van.dolder@vu.nl; [e] Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom, d.vandolder@essex.ac.uk; [f] Institute for Advanced Economic Research, Dongbei University of Finance and Economics, 217 Jianshan Street, Dalian, Liaoning 116025, P.R. China, t.wang@dufe.edu.cn

Students do not attend all courses for which they register. Therefore, in the paper, we restricted our sample to the set of courses for which the student obtained a final grade (which could be a failing grade) rather than all courses for which the student registered. Students who did not obtain a final grade for a course were unlikely to evaluate that course. They only did so 10 percent of the time compared to 23 percent for students who did obtain a final grade. As a robustness check, this online Appendix presents the analyses when we consider all courses for which a student registered.

Now that we consider all courses for which a student registered, we define course size as the number of students who registered for the course, instead of the number who got a final grade as we did in the paper. The average grade of a student is the average of all the grades obtained (omitting the courses where she registered but did not get a final grade) and could be a missing value. Course grade (deviation from average) is a missing value for courses for which the student did not obtain a grade.

Table OA.1 shows the summary statistics. The average grade of a student deciding to complete a survey was 6.64 (out of 10). Course size varied between 3 and 864 students. Furthermore, 23 percent of the observations pertained to master-level courses, while 77 percent pertained to bachelor-level courses. Tests of equality of variable means across treatments provide no statistically significant evidence for imbalances in these observable characteristics (Average grade: clustered $F(3,4567) = 1.69,\ p = 0.167$; Course grade (deviation from average): clustered $F(3,4567) = 1.19,\ p = 0.311$; Course size: clustered $F(3,4805) = 2.00,\ p = 0.111$; Master: clustered $Chi^2(3) = 1.935,\ p = 0.586$). The largest absolute value of the normalized differences that we observe is below 0.1, lower than the 0.25 threshold suggested by Imbens and Rubin (2015), suggesting that simple regression methods are sufficient to control for imbalances in covariates in our data.

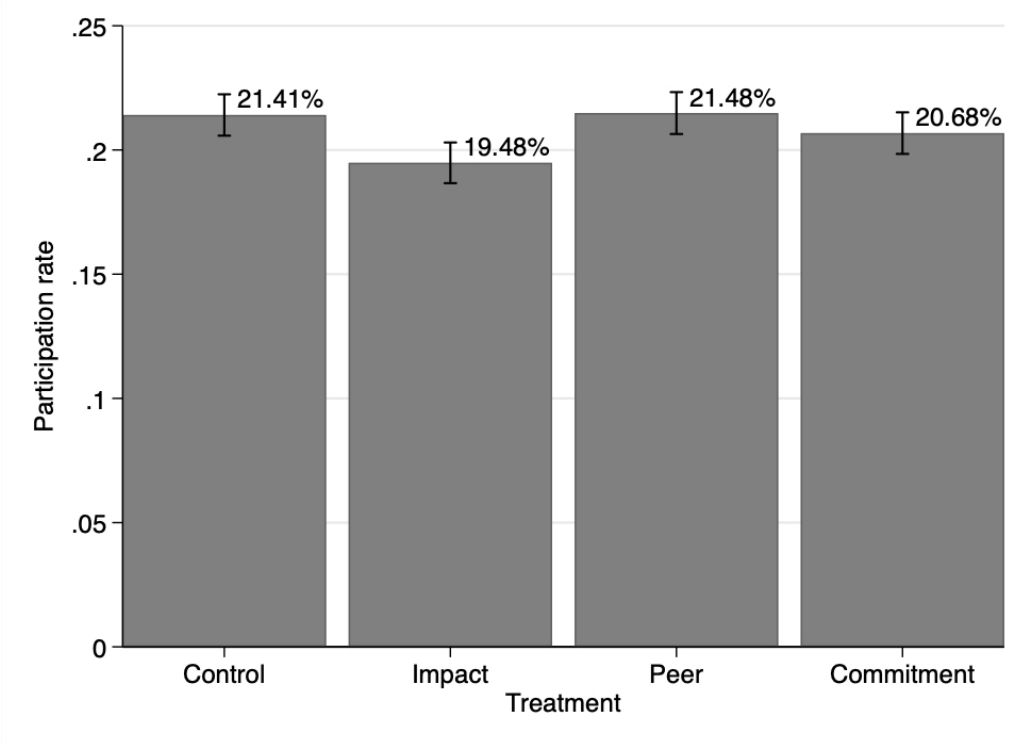
Table OA.1. Summary statistics

	Total			Mean (std. dev.) [number of missing values, if any] by treatment			
	Mean (std. dev.) [number of missing values, if any]	Min	Max	Control	Impact	Peer	Commit-ment
Average grade	6.64 (1.24) [573]	1.00	9.84	6.69 (1.25) [182]	6.59 (1.23) [148]	6.68 (1.21) [134]	6.60 (1.25) [109]
Course grade	0 (1.09) [6,242]	-6.65	5.08	0 (1.06) [1,572]	0 (1.14) [1,571]	0 (1.07) [1,598]	0 (1.09) [1,501]
Course size	301 (232)	3	864	305 (232)	302 (233)	289 (225)	309 (237)
Master	0.23 (0.42)	0	1	0.23 (0.42)	0.22 (0.41)	0.24 (0.43)	0.24 (0.42)
Number of observations	36,463			9,347	9,003	9,118	8,995

Figure OA.1 shows the students' rates of participation in the course evaluations across the four treatments. The participation rates in the four treatments were 21.41% (control), 19.48% (impact), 21.48% (peer), and 20.68% (commitment). In the commitment treatment, 11.54% of the students responded to the commitment email, of which 90.79% committed with a "yes". Participation rates did not differ significantly across the four treatments (clustered $Chi^2(3) = 2.984, p = 0.394$). In binary comparisons, none of the treatments differed significantly from the control treatment (impact vs. control: clustered $Chi^2(1) = 2.174, p = 0.140$; peer vs. control: clustered $Chi^2(1) = 0.003, p = 0.954$; commitment vs. control: clustered $Chi^2(1) = 0.313, p = 0.576$). Directionally, the participation rates in both the impact treatment and the commitment treatment were even a bit lower than that in the control treatment.

Table OA.2 presents the result of logistic regression analyses, where the dependent variable is a dummy variable that takes the value one if the student participated in the course evaluation and zero otherwise. As in the paper, we correct standard errors for clustering at the student level.

Figure OA.1. Participation rates by treatment



Note: The confidence bars represent the 95% confidence intervals.

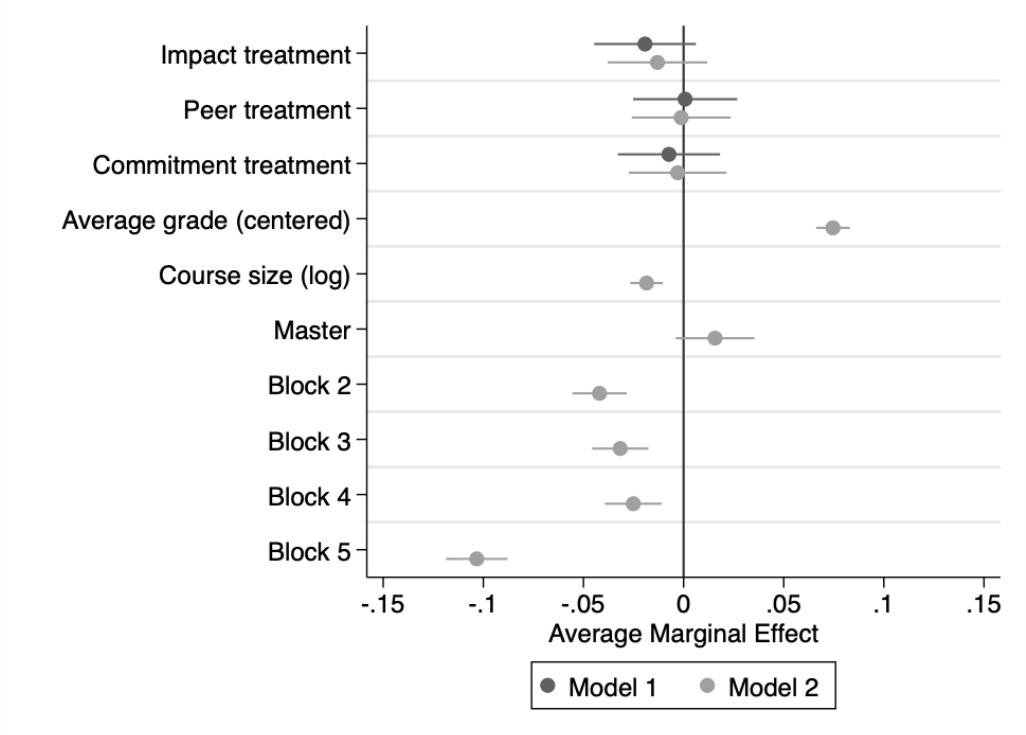
Table OA.2, Model 1 only includes the treatment dummies and confirms that none of our nudges significantly affected student participation in course evaluations (all $p > 0.136$). Including course grades in the regression would force all the observations to have a valid grade, which would lead to the same sample as in the paper. Therefore, Table OA.2, Model 2 only includes controls for the student's average grade (centered on the student-level average 6.54), the course size (log-transformed), the course level (bachelor vs. master), and the block in which the course took place. Again, we find insignificant treatment effects for all nudges (all $p > 0.301$). In addition, we find that students with better average grades were more likely to complete course evaluations ($p < 0.001$). Also, students were less likely to complete evaluations for larger courses ($p < 0.001$). Master-level students did not differ significantly from bachelor-level students in their likelihood to evaluate courses ($p = 0.124$). Participation rates were highest in the first block of the year and lowest in the fifth (last) block.

Table OA.2. Logit regression results

	Dependent variable: Completed (1 if the student completed course evaluation, 0 otherwise)		
	Model 1 coeff (p-value)	Model 2 coeff (p-value)	Model 3 coeff (p-value)
Impact	-0.118 (0.137)	-0.084 (0.302)	-0.153 (0.710)
Peer	0.005 (0.954)	-0.007 (0.925)	0.436 (0.280)
Commitment	-0.044 (0.575)	-0.019 (0.813)	0.241 (0.556)
Average grade (centered)		0.477 (0.000)	0.384 (0.000)
Course size (log)		-0.117 (0.000)	-0.066 (0.209)
Master (1 if master-level)		0.097 (0.124)	-0.069 (0.572)
Average grade * impact			0.040 (0.609)
Average grade * peer			0.137 (0.062)
Average grade * commitment			0.213 (0.006)
Course size * impact			-0.012 (0.874)
Course size * peer			-0.109 (0.136)
Course size * commitment			-0.073 (0.323)
Master * impact			0.401 (0.022)
Master * peer			0.200 (0.248)
Master * commitment			0.077 (0.658)
Block 2		-0.259 (0.000)	-0.259 (0.000)
Block 3		-0.193 (0.000)	-0.191 (0.000)
Block 4		-0.150 (0.001)	-0.149 (0.001)
Block 5		-0.714 (0.000)	-0.714 (0.000)
Constant	-1.301 (0.000)	-0.647 (0.000)	-0.823 (0.005)
Level of clustering	student	student	student
Number of observations	36,463	35,890	35,890
McFadden pseudo R-squared	0.0004	0.0579	0.0601

Figure OA.2 displays the average marginal treatment effects implied by Model 1 and Model 2. As shown in Figure OA.2, the marginal effects of our nudge treatments on students' likelihood of participation are close to zero across the board. By contrast, the marginal effects estimated for changes in the students' average grades, the course sizes, and the academic blocks are non-zero and tend to be larger in magnitude than the treatment effects.

Figure OA.2. Average marginal effects implied by Model 1 and Model 2

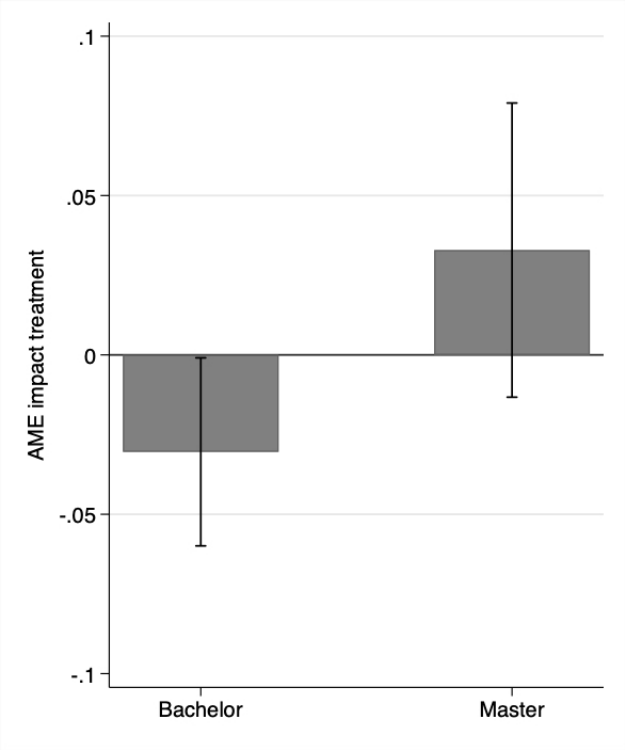


Note: Nodes show the average marginal effects on students' likelihood of participation in the course evaluations. Lines indicate the 95% confidence intervals for the marginal effects.

Table OA.2, Model 3 includes interactions of each nudge treatment with the average grade, the course size, and the course level (master vs. bachelor). Overall, adding these nine interactions significantly improves model performance ($Chi^2(9) = 20.26, p = 0.016$). Exploring interactions in more detail, we observe that course size did not significantly interact with any of the treatments (all $p > 0.135$). However, master-level students appear to have been more positively influenced by the impact treatment than students attending bachelor-level courses ($p = 0.022$). Finally, there is evidence that the commitment treatment was more effective for students with higher average grades ($p = 0.006$). Both the interaction of the impact treatment with course level and the interaction of the commitment treatment with average grade remain significant after applying the Romano-Wolf multiple-hypothesis correction (Clarke et al. 2020; see also Romano et al. 2010). Figures OA.3 and OA.4 illustrate these interactions.

Figure OA.3 plots the estimated average effects of the impact treatment for bachelor vs. master-level courses. We observe that the estimated effect size is negative for students attending bachelor-level courses but positive for students attending master-level courses. However, only the negative effect for bachelor students is statistically significant; see exact estimates in Table OA.B1.

Figure OA.3. Impact treatment effects for bachelor-level and master-level courses (Model 3 estimates)



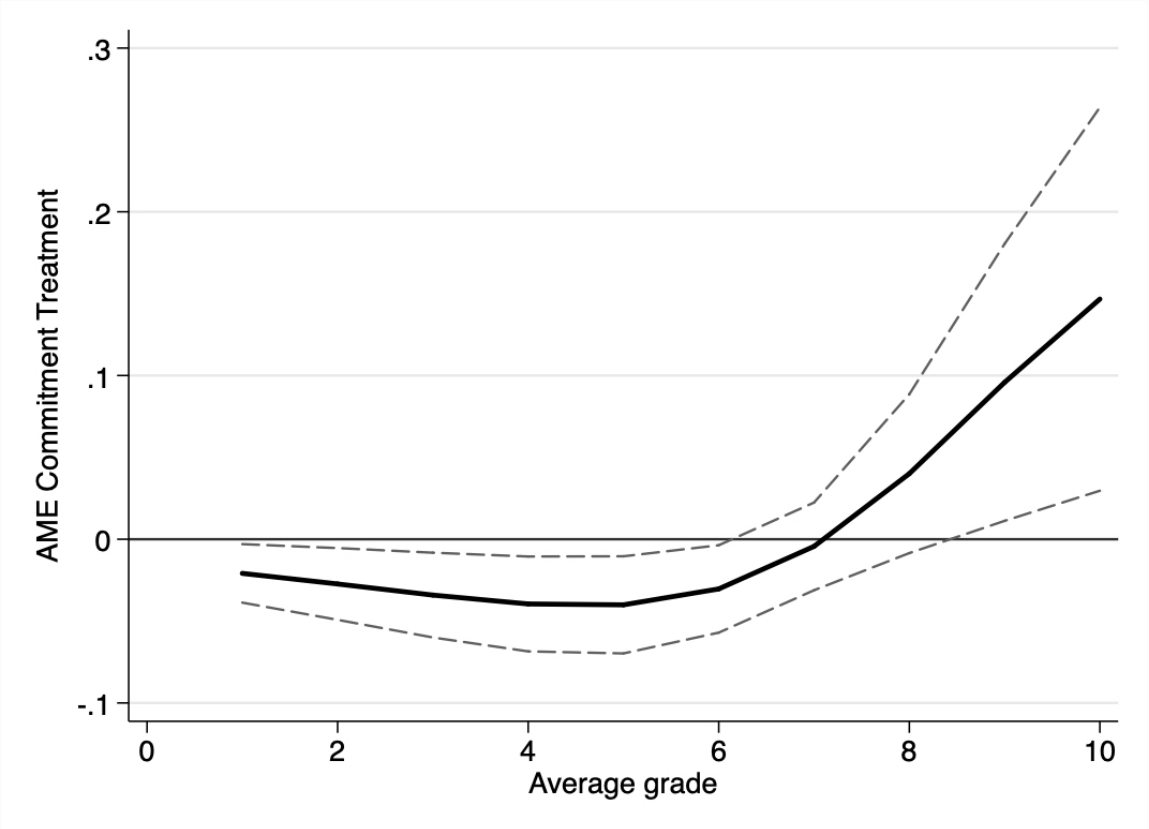
Note: The confidence bars represent the 95% confidence intervals. Estimates are obtained from Model 3 of Table 2.

Table OA.B1. Impact treatment effects (on the likelihood of participation) as a function of course level

		Avg. marginal effect in percentage points (p-value)
Master	0	-3.064 (0.042)
	1	3.344 (0.157)

Figure OA.4 plots the estimated average effects of the commitment treatment as a function of the student's average grade. We observe that the commitment treatment increased the predicted likelihood of participation of students with good average grades. By comparison, the commitment treatment decreased the participation likelihood of students with poor average grades; see exact estimates in Table OA.B2.

Figure OA.4. Commitment treatment effects as a function of average grade (Model 3 estimates)



Note: Solid line plots the estimated marginal effects of the commitment treatment dummy. Dashed lines indicate the 95% confidence intervals. Estimates are obtained from Model 3 of Table OA.2.

Table OA.B2. Commitment treatment effects (on the likelihood of participation) as a function of average grade

		Avg. marginal effect in percentage points (p-value)
Average grade	1	-2.088 (0.022)
	2	-2.734 (0.014)
	3	-3.417 (0.010)
	4	-3.958 (0.007)
	5	-4.013 (0.008)
	6	-3.040 (0.026)
	7	-0.438 (0.748)
	8	4.011 (0.105)
	9	9.596 (0.026)
	10	14.696 (0.014)

Table OA.B3. Logit regression results for models exploring the mechanisms within the commitment treatment

	Dependent variable			
coeff (p-value)	Responded to commitment email	Responded with "yes" (if responded)	Completed evaluation (if responded with "yes")	Completed evaluation (if did not respond to commitment email)
Average grade (centered)	0.564 (0.000)	0.339 (0.010)	0.440 (0.000)	0.488 (0.000)
Course size (log)	0.003 (0.965)	-0.315 (0.131)	-0.135 (0.243)	-0.188 (0.000)
Master (1 if master-level)	-0.277 (0.114)	0.265 (0.591)	0.016 (0.956)	0.219 (0.095)
Block 2	-0.417 (0.000)	0.340 (0.340)	0.265 (0.288)	-0.366 (0.007)
Block 3	-0.744 (0.000)	0.306 (0.551)	1.350 (0.000)	0.087 (0.459)
Block 4	-0.793 (0.000)	0.113 (0.811)	1.124 (0.002)	0.121 (0.305)
Block 5	-1.332 (0.000)	0.701 (0.421)	0.859 (0.143)	-0.308 (0.049)
Constant	-1.574 (0.001)	3.600 (0.004)	1.420 (0.034)	-1.028 (0.001)
Level of clustering	student	student	student	student
Number of observations	8,886	1,110	1,007	7,776
McFadden pseudo R-squared	0.0709	0.0514	0.0860	0.0630

Exploring the commitment treatment in more detail, we observe that students with higher average grades responded more positively at each stage of the commitment treatment process. Compared to students with poor average grades, students with high grades were: (i) more likely to respond to the commitment email; (ii) conditionally on responding, more likely to respond in the affirmative; and (3) more likely to meet their commitment and actually complete the course evaluation after they had committed; see Table OA.B3.