European Economic Review 141 (2022) 104001

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

Nudging student participation in online evaluations of teaching: Evidence from a field experiment[1]

Susanne Neckermann [a], Uyanga Turmunkh [b,*], Dennie van Dolder [c,d], Tong V. Wang [e]

[a] Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, United States
[b] IESEG School of Management, Univ. Lille, CNRS, UMR 9221 - LEM - Lille Economie Management, F-59000 Lille, France
[c] School of Business and Economics, Vrije Universiteit (VU) Amsterdam, Gustav Mahlerplein 117, 1082 MS Amsterdam, the Netherlands
[d] Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom
[e] Institute for Advanced Economic Research, Dongbei University of Finance and Economics, 217 Jianshan Street, Dalian, Liaoning 116025, China

ARTICLE INFO

JEL classification:
D90
D91
H41
I20

Keywords:
Nudges
Social norms
Descriptive norm
Commitment
Student evaluation of teaching
Participation
Response rates
Field experiment

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.

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

[1] We thank the Education Service Center at Erasmus University Rotterdam and in particular Ria Koolen-Verhulst, Guus Eilers, Harry Post, Robert Kwakkel, Remco Slim, Tom van Steenbrugge, and Nabil el Issati for invaluable assistance in conducting the experiment and for providing us with the data. Furthermore, we thank Martijn van den Assem, Robert Dur, Ferdinand Vieider, and the two anonymous reviewers for their constructive comments. The paper has benefitted from discussion with seminar participants at Erasmus University Rotterdam.

* Corresponding author.
E-mail addresses: sneckermann@uchicago.edu (S. Neckermann), u.turmunkh@ieseg.fr (U. Turmunkh), d.van.dolder@vu.nl, d.vandolder@essex.ac.uk (D. van Dolder), t.wang@dufe.edu.cn (T.V. Wang).

https://doi.org/10.1016/j.euroecorev.2021.104001
Received 11 August 2020; Received in revised form 12 November 2021; Accepted 26 November 2021
Available online 9 December 2021
0014-2921/

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.

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 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

[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 fines (National Science and Technology Council 2016, The Behavioural Insights Team 2016, The Behavioural Insights Team 2017, The Behavioural Insights Team 2019).

[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.

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 experiment included all 3485 bachelor and 1552 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 fund-raising 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 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

[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.

[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.

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; 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

Fig. 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.

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

[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			7775	7432	7520	7494

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.

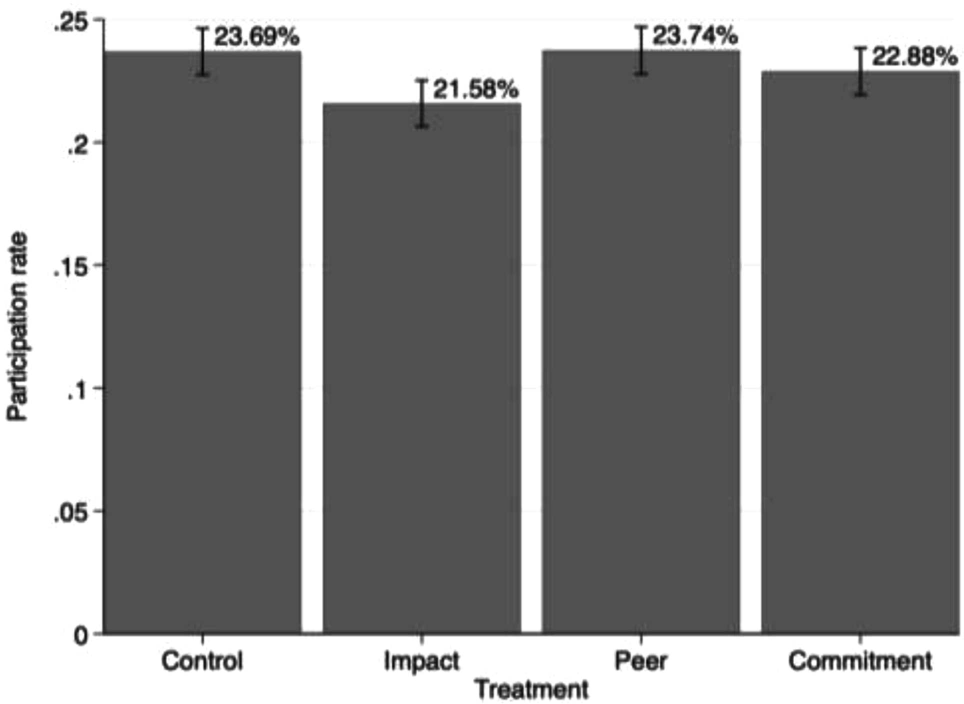

Fig. 1. Participation rates by treatment
Note: The confidence bars represent the 95% confidence intervals.

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.

Because the logistic regression model coefficients are not easy to interpret, Fig. 2 displays the average marginal treatment effects implied by Models 1 and 2. As shown in Fig. 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.

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

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

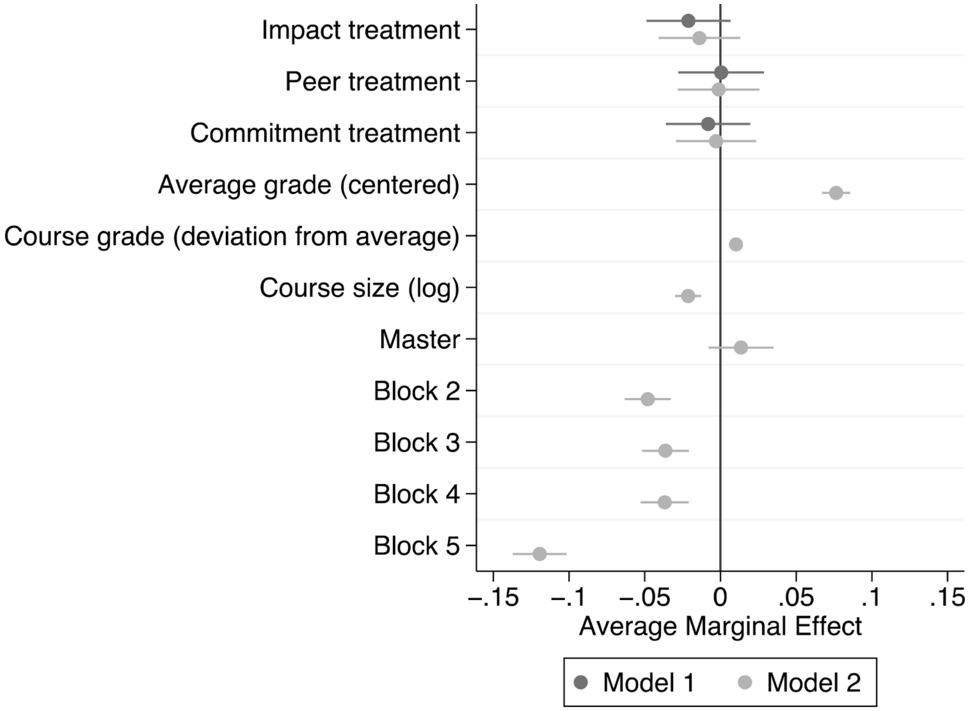

Fig. 2. Average marginal effects implied by Model 1 and Model 2
Note: Nodes show the average marginal effects on students' likelihood of participation in the course evaluations. Lines indicate the 95% confidence intervals for the marginal effects.

($Chi^2(12) = 24.86,\ p = 0.016$). 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). Figs. 3 and 4 illustrate these interactions.

Fig. 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).

Fig. 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).

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

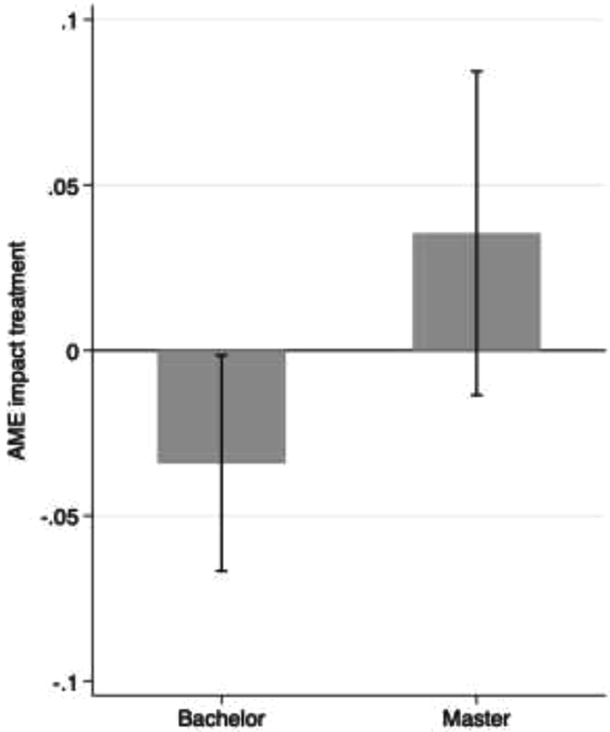

Fig. 3. Impact treatment effects for bachelor-level and master-level courses (Model 3 estimates)
Note: The confidence bars represent the 95% confidence intervals. Estimates are obtained from Model 3 of Table 2.

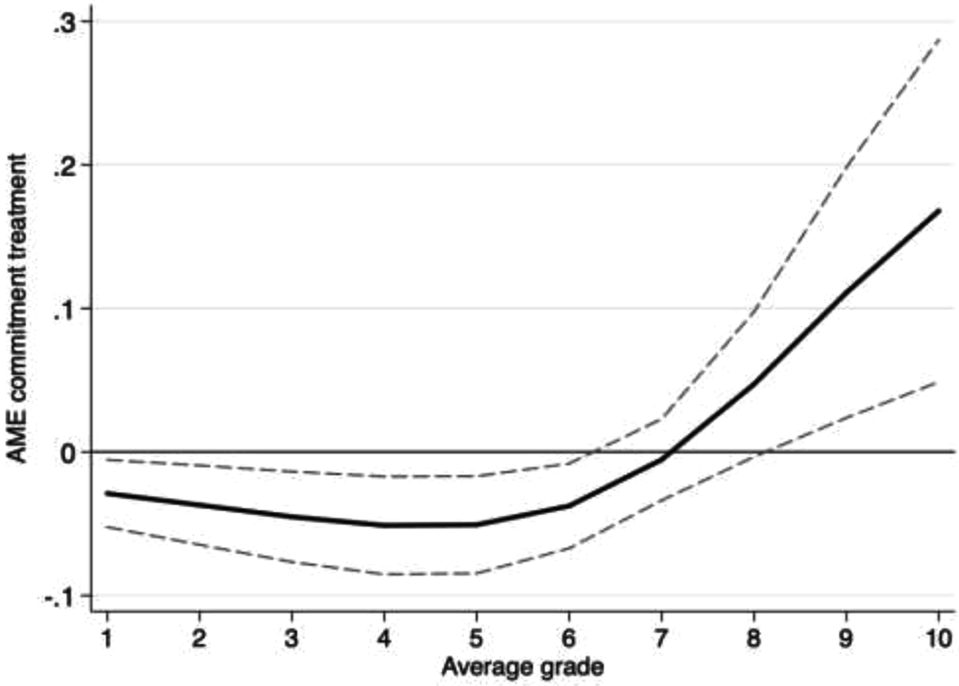

Fig. 4. Commitment treatment effects as a function of average grade (Model 3 estimates)
Note: Solid line plots the estimated marginal effects of the commitment treatment dummy. Dashed lines indicate the 95% confidence intervals. Estimates are obtained from Model 3 of Table 2.

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)

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

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

coeff (p-value)	*Dependent variable* 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	0.261	0.423	0.490
	(0.000)	(0.072)	(0.001)	(0.000)
Course grade (deviation from average grade)	-0.007	0.092	-0.067	0.128
	(0.750)	(0.409)	(0.426)	(0.000)
Course size (log)	-0.004	-0.381	-0.144	-0.208
	(0.957)	(0.090)	(0.258)	(0.000)
Master (1 if master-level)	-0.332	0.289	0.079	0.187
	(0.062)	(0.573)	(0.789)	(0.169)
Block 2	-0.432	0.291	0.273	-0.410
	(0.000)	(0.431)	(0.305)	(0.003)
Block 3	-0.775	0.625	1.379	0.089
	(0.000)	(0.201)	(0.000)	(0.460)
Block 4	-0.849	0.171	1.450	0.039
	(0.000)	(0.743)	(0.001)	(0.748)
Block 5	-1.265	0.426	1.022	-0.410
	(0.000)	(0.628)	(0.134)	(0.020)
Constant	-1.463	3.962	1.498	-0.858
	(0.002)	(0.003)	(0.034)	(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

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.

Declaration of Competing Interest

Our field experiment was run with the help from Education Service Center of Erasmus School of Economics, which also subsequently provided us access to relevant data in anonymous format. The Education Service Center, however, played no role in study design, analysis, preparation of manuscript, or publication decision. No other third-party played any role. Disclosure statement for Susanne Neckermann: I have no conflicts to disclose on this paper. I received no funding and have no financial interest. Susanne NeckermannUyanga Turmunkh has no conflicts of interest to disclose. The author received no funding for this research. Dennie van Dolder gratefully acknowledge support from the Netherlands organization for Scientific Research (NWO, 452–16–011) and from the Economic and Social Research Council via the Network for Integrated Behavioural Sciences (ES/K002201/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Tong V. Wang declares that she has no material financial interests that relate to the research described in this paper.

Appendix B. Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2021.104001.

References

Adams, Meredith JD, Umbach, Paul D., 2012. Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. Research in Higher Education 53 (5), 576–591. https://doi.org/10.1007/s11162-011-9240-5.

Allcott, Hunt., 2011. Social norms and energy conservation. J. Public Econ. 95 (9–10), 1082–1095. https://doi.org/10.1016/j.jpubeco.2011.03.003.

Allcott, Hunt, Rogers, Todd, 2014. The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. Am. Econ. Rev. 104 (10), 3003–3037. https://doi.org/10.1257/aer.104.10.3003.

Alter, Molly, Reback, Randall, 2014. True for your school? How changing reputations alter demand for selective US colleges. Educational Evaluation and Policy Analysis 36 (3), 346–370. https://doi.org/10.3102/0162373713517934.

Alvero, Alicia M., Mangiapanello, Kathleen, Valad, Jennifer, 2019. 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 (4), 501–515. https://doi.org/10.1080/02602938.2018.1521913.

Ariely, Dan, Kamenica, Emir, Prelec, Dražen, 2008. Man's search for meaning: The case of Legos. J. Econ. Behav. Org. 67 (3–4), 671–677. https://doi.org/10.1016/j.jebo.2008.01.004.

Aronson, Elliot., 1992. The return of the repressed: Dissonance theory makes a comeback. Psychological Inquiry 3 (4), 303–311. https://doi.org/10.1207/s15327965pli0304_1.

Avery, Rosemary J., Bryant, W.Keith, Mathios, Alan, Kang, Hyojin, Bell, Duncan, 2006. Electronic course evaluations: does an online delivery system influence student evaluations? J. Econ. Edu. 37 (1), 21–37. https://doi.org/10.3200/JECE.37.1.21-37.

Ayres, Ian, Raseman, Sophie, Shih, Alice, 2013. Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. J. Law Econ. Org. 29 (5), 992–1022. https://doi.org/10.1093/jleo/ews020.

Baca-Motes, Katie, Brown, Amber, Gneezy, Ayelet, Keenan, Elizabeth A., Nelson, Leif D., 2013. Commitment and behavior change: Evidence from the field. J. Consum. Res. 39 (5), 1070–1084. https://doi.org/10.1086/667226.

Bartke, Simon, Friedl, Andreas, Gelhaar, Felix, Reh, Laura, 2017. Social comparison nudges—Guessing the norm increases charitable giving. Economics Letters 152, 73–75. https://doi.org/10.1016/j.econlet.2016.12.023.

Becker, William E., Bosshardt, William, Watts, Michael, 2012. How departments of economics evaluate teaching. J. Econ. Edu. 43 (3), 325–333. https://doi.org/10.1080/00220485.2012.686826.

Becker, William E., Watts, Michael, 1999. How departments of economics evaluate teaching. Am. Econ. Rev. 89 (2), 344–349. https://doi.org/10.1257/aer.89.2.344.

Bernstein, Edward, Edwards, Erika, Dorfman, David, Heeren, Tim, Bliss, Caleb, Bernstein, Judith, 2009. Screening and brief intervention to reduce marijuana use among youth and young adults in a pediatric emergency department. Acad. Emerg. Med. 16 (11), 1174–1185. https://doi.org/10.1111/j.1553-2712.2009.00490.x.

Bott, Kristina M., Cappelen, Alexander W., Sørensen, Erik Ø., Tungodden, Bertil, 2020. You've got mail: A randomized field experiment on tax evasion. Management Science 66 (7). https://doi.org/10.1287/mnsc.2019.3390.

Chabé-Ferret, Sylvain, Coent, Philippe Le, Reynaud, Arnaud, Subervie, Julie, Lepercq, Daniel, 2019. Can we nudge farmers into saving water? Evidence from a randomised experiment. European Review of Agricultural Economics 46 (3), 393–416. https://doi.org/10.1093/erae/jbz022.

Chapman, Diane D., Joines, Jeffrey A., 2017. Strategies for increasing response rates for online end-of-course evaluations. Int. J. Teach. Learn. High. Edu. 29 (1), 47–60. Available online.

Chen, Yining, Hoshower, Leon B., 2003. Student evaluation of teaching effectiveness: An assessment of student perception and motivation. Assessment and Evaluation in Higher Education 28 (1), 71–88. https://doi.org/10.1080/02602930301683.

Chong, Alberto, Karlan, Dean, Shapiro, Jeremy, Zinman, Jonathan, 2015. (Ineffective) messages to encourage recycling: evidence from a randomized evaluation in Peru. The World Bank Economic Review 29 (1), 180–206. https://doi.org/10.1093/wber/lht022.

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

Cialdini, Robert B., Reno, Raymond R., Kallgren, Carl A., 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. J. Pers. Soc. Psychol. 58 (6), 1015. https://doi.org/10.1037/0022-3514.58.6.1015.

Clarke, Damian, Romano, Joseph P., Wolf, Michael, 2020. The Romano–Wolf multiple-hypothesis correction in Stata. The Stata J. 20 (4), 812–843. https://doi.org/10.1177/1536867X20976314.

Costa, Dora L., Kahn, Matthew E., 2013. Energy conservation 'nudges' and environmentalist ideology: Evidence from a randomized residential electricity field experiment. J. Eur. Econ. Assoc. 11 (3), 680–702. https://doi.org/10.1111/jeea.12011.

DellaVigna, Stefano, Linos, Elizabeth, 2020. RCTs to scale: Comprehensive evidence from Two Nudge Units. NBER Working Paper, p. 27594. Available online.

Dommeyer, Curt J., Baum, Paul, Hanna, Robert W., Chapman, Kenneth S., 2004. Gathering Faculty Teaching Evaluations by in-class and Online Surveys: Their Effects on Response Rates and Evaluations. Assessment and Evaluation in Higher Education 29 (5), 611–623. https://doi.org/10.1080/02602930410001689171.

Farrow, Katherine, Grolleau, Gilles, Ibanez, Lisette, 2017. Social norms and pro-environmental behavior: A review of the evidence. Ecol. Econ. 140, 1–13. https://doi.org/10.1016/j.ecolecon.2017.04.017.

Festinger, Leon., 1957. A Theory of Cognitive Dissonance, 2. Stanford university press.

Gerber, Alan S., Rogers, Todd, 2009. Descriptive social norms and motivation to vote: Everybody's voting and so should you. J. Pol. 71 (1), 178–191. https://doi.org/10.1017/S0022381608090117.

Goldstein, Noah J., Cialdini, Robert B., Griskevicius, Vladas, 2008. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. J. Consum. Res. 35 (3), 472–482. https://doi.org/10.1086/586910.

Goodman, James, Anson, Robert, Belcheir, Marcia, 2015. The effect of incentives and other instructor-driven strategies to increase online student evaluation response rates. Assessment and Evaluation in Higher Education 40 (7), 958–970. https://doi.org/10.1080/02602938.2014.960364.

Goos, Maarten, Salomons, Anna, 2017. Measuring teaching quality in higher education: assessing selection bias in course evaluations. Research in Higher Education 58 (4), 341–364. https://doi.org/10.1007/s11162-016-9429-8.

Grant, Adam M., 2008. Employees without a cause: The motivational effects of prosocial impact in public service. Int. Public Manage. J. 11 (1), 48–66. https://doi.org/10.1080/10967490801887905.

Greenwald, Anthony G., Carnot, Catherine G., Beach, Rebecca, Young, Barbara, 1987. Increasing voting behavior by asking people if they expect to vote. J. Appl. Psychol. 72 (2), 315. https://doi.org/10.1037/0021-9010.72.2.315.

Hallsworth, Michael, List, John A., Metcalfe, Robert D., Vlaev, Ivo, 2017. The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. J. Public Econ. 148, 14–31. https://doi.org/10.1016/j.jpubeco.2017.02.003.

Hummel, Dennis, Maedche, Alexander, 2019. How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. J. Behav. Exp. Econ. 80, 47–58. https://doi.org/10.1016/j.socec.2019.03.005.

Imbens, Guido W., Rubin, Donald B., 2015. Causal Inference in statistics, social, and Biomedical Sciences. Cambridge University Press.

Johnson, Rachel., 2000. The authority of the student evaluation questionnaire. Teaching in Higher Education 5 (4), 419–434. https://doi.org/10.1080/713699176.

Johnson, Trav D., 2003. Online Student Ratings: Will Students Respond? New Directions for Teaching and Learning 96, 49–59. https://doi.org/10.1002/tl.122.

Kettle, Stewart, Hernandez, Marco, Ruda, Simon, Sanders, Michael, 2016. Behavioral Interventions in Tax compliance: Evidence from Guatemala. The World Bank. Available online.

Kherfi, Samer., 2011. Whose opinion is it anyway? Determinants of participation in student evaluation of teaching. J. Econ. Edu. 42 (1), 19–30. https://doi.org/10.1080/00220485.2011.536487.

Kosfeld, Michael, Neckermann, Susanne, Yang, Xiaolan, 2017. The effects of financial and recognition incentives across work contexts: The role of meaning. Econ. Inq. 55 (1), 237–247. https://doi.org/10.1111/ecin.12350.

Layne, Benjamin H., DeCristoforo, Joseph R., McGinty, Dixie, 1999. Electronic versus traditional student ratings of instruction. Research in Higher Education 40 (2), 221–232. https://doi.org/10.1023/A:1018738731032.

Li, Xueheng, Molleman, Lucas, van Dolder, Dennie, 2021. Do descriptive social norms drive peer punishment? Conditional punishment strategies and their impact on cooperation. Evol. Hum. Behav. 42 (5), 469–479. https://doi.org/10.1016/j.evolhumbehav.2021.04.002.

Lipsey, Nikolette, Shepperd, James, 2020. Examining strategies to increase student evaluation of teaching completion rates. Assessment and Evaluation in Higher Education. https://doi.org/10.1080/02602938.2020.1782343.

Mann, Christopher B., 2005. Unintentional voter mobilization: Does participation in preelection surveys increase voter turnout? Ann. Am. Acad. Pol. Soc. Sci. 601 (1), 155–168. https://doi.org/10.1177/0002716205278151.

Misra, Shalini, Stokols, Daniel, Marino, Anne Heberger, 2012. Using norm–based appeals to increase response rates in evaluation research: A field experiment. Am. J. Eval. 33 (1), 88–98. https://doi.org/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., Rogers, Todd, 2010. Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making. Psychol. Sci. 21 (2), 194–199. https://doi.org/10.1177/0956797609359326.

Nolan, Jessica M., Schultz, P.Wesley, Cialdini, Robert B., Goldstein, Noah J., Griskevicius, Vladas, 2008. Normative social influence is underdetected. Pers. Soc. Psychol. Bull. 34 (7), 913–923. https://doi.org/10.1177/0146167208316691.

Nulty, Duncan D., 2008. The adequacy of response rates to online and paper surveys: what can be done?". Assessment and Evaluation in Higher Education 33 (3), 301–314. https://doi.org/10.1080/02602930701293231.

Parks, Craig D., Sanna, Lawrence J., Berel, Susan R., 2001. Actions of similar others as inducements to cooperate in social dilemmas. Pers. Soc. Psychol. Bull. 27 (3), 345–354. https://doi.org/10.1177/0146167201273008.

Perkins, H.Wesley, 2003. 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 3–17.

Reisenwitz, Timothy H., 2016. Student evaluation of teaching: An investigation of nonresponse bias in an online context. J. Market. Edu. 38 (1), 7–17. https://doi.org/10.1177/0273475315596778.

Romano, Joseph P., Shaikh, Azeem M., Wolf, Michael, 2010. Hypothesis testing in econometrics. Ann. Rev. Econ. 2 (1), 75–104. https://doi.org/10.1146/annurev.economics.102308.124342.

Sandberg, Tracy, Conner, Mark, 2009. A mere measurement effect for anticipated regret: Impacts on cervical screening attendance. Br. J. Soc. Psychol. 48 (2), 221–236. https://doi.org/10.1348/014466608X347001.

Sanders, Michael, Snijders, Veerle, Hallsworth, Michael, 2018. Behavioural science and policy: where are we now and where are we going? Behavioural Public Policy 2 (2), 144–167. https://doi.org/10.1017/bpp.2018.17.

Schultz, Wesley P., Khazian, Azar M., Zaleski, Adam C., 2008. Using normative social influence to promote conservation among hotel guests. Social Influence 3 (1), 4–23. https://doi.org/10.1080/15534510701755614.

Shang, Jen, Croson, Rachel, 2009. A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. Econ. J. 119 (540), 1422–1439. https://doi.org/10.1111/j.1468-0297.2009.02267.x.

Smith, Jennifer K., Gerber, Alan S., Orlich, Anton, 2003. Self-prophecy effects and voter turnout: An experimental replication. Political Psychology 24 (3), 593–604. https://doi.org/10.1111/0162-895X.00342.

Spencer, Karin J., Schmelkin, Liora Pedhazur, 2002. Student perspectives on teaching and its evaluation. Assessment and Evaluation in Higher Education 27 (5), 397–409. https://doi.org/10.1080/0260293022000009285.

Spooren, Pieter, Brockx, Bert, Mortelmans, Dimitri, 2013. On the validity of student evaluation of teaching: The state of the art. Review of Educational Research 83 (4), 598–642. https://doi.org/10.3102/0034654313496870.

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

Sundstrom, Eric D., Hardin, Erin E., Shaffer, Matthew J., 2016. Extra credit micro-incentives and response rates for online course evaluations: Two quasi-experiments. Teaching of Psychology 43 (4), 276–284. https://doi.org/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, Dahlstrand, Ulf, Biel, Anders, 1999. From Ought to Is: Moral Norms in Large-Scale Social Dilemmas. Goteborg Psychological Reports 29 (5), 1–17. Available online.
Wilding, Sarah, Conner, Mark, Sandberg, Tracy, Prestwich, Andrew, Lawton, Rebecca, Wood, Chantelle, Miles, Eleanor, Godin, Gaston, Sheeran, Paschal, 2016. The question-behaviour effect: a theoretical and methodological review and meta-analysis. European Review of Social Psychology 27 (1), 196–230. https://doi.org/10.1080/10463283.2016.1245940.