The power of (no) recognition: Experimental evidence from the university classroom

Nicky Hoogveld and Nick Zubanov*

Abstract

We study the effect of unannounced recognition on performance with a field experiment involving first-year Dutch university students attending tutorials as part of a compulsory course. Our treatment, given in randomly selected tutorial groups, was to publicly recognize students who scored within the top 30% of their respective group on the first of the two midterm exams. The overall treatment effect on the second midterm grade is $0.03s$ (s = grade standard deviation) for the recipients of recognition, and $0.15s$ for the non-recipients, both statistically insignificant. The effect for the non-recipients increases with class attendance (itself unaffected by our treatment) and proximity to the cutoff grade for recognition, reaching a significant $0.55s$ for the 23% of the non-recipients who attended at least 12 out of 13 classes and were within the first quartile of the distance to cutoff. We argue that conformity to performance norm is among the forces shaping the effects we observe.

*Corresponding author. University of Konstanz, `nick.zubanov@uni-konstanz.de`

1 Introduction

Recognition is one of the core practices in education. Our study presents experimental evidence on the effects of this practice on university student performance. There are (at least) three reasons why economists should be interested in the effect of recognition on student performance. *First*, because it is cheap, recognition may be a more efficient alternative to financial incentives for students and teachers, class size reduction, or extra academic support. Indeed, as Levitt et al. (2016) found from a series of experiments with Chicago school students, a symbolic award – a trophy and a photo on the wall in the class, costing about $3 – improved grade 2 to 5 students' test score by 0.12 of its standard deviation, on a par with the effect of financial incentives of up to $20.

Second, depending on how it is provided, recognition can affect not only average performance but also performance distribution. For instance, in Bradler et al.'s (2016) experiment, where recognition was given unannounced, it had a bigger effect on non-recipients' than recipients' output in a data entry task (0.5 of standard deviation for non-recipients vs. 0.2 for recipients). Chen et al. (2010), who provided unannounced performance feedback, found a similar effect: a decrease in output by those performing above the median, and a large increase by those below. On the other hand, when recognition was announced, as in Kosfeld and Neckermann's (2011) experiment, it triggered a higher response from the more able. Thus, when the goal is to improve the performance of their currently underachieving students, unannounced recognition could be a solution.

Third, the university is a relatively tough environment for the effect of recognition to be felt, because there are other, powerful and universally applicable, reasons for students to do well, such as passing the course and, ultimately, graduating. Whether recognition continues to affect performance in this environment is, therefore, an open question. In fact, the positive effect of recognition for grade 2-5 students in Levitt et al. (2016) declines into insignificance for more senior, grade 6-8 students, precisely when those reasons become more important. Carried out in an environment where powerful incentives other than recognition are present, our study differs from the existing literature on the subject much of which excludes these incentives by designing one-off jobs with fixed pay and no career concerns (Kosfeld and Neckermann, 2011; Kube, Marechal and Puppe, 2012; Bradler et al., 2016).

In Harrison and List's (2004) classification, ours is a "natural field experiment", run in an environment that was perfectly normal for the participants, and without them knowing they were part of an experiment. Our sample consists of 368 first-year undergraduate students at a Dutch university attending the compulsory microeconomics course, of whom 342 have a complete grade record. Everyone had to pass this course in order to continue with their studies; hence the presence of strong, uniform incentives to do well. Before the start of the academic year, and without our involvement, these students were randomly divided into 15 tutorial groups each taught by an experienced teaching assistant (TA). Our treatment took place between the two midterm exams during the course, each carrying an equal weight in the final grade (10%). The treatment, administered through and on behalf of the TAs in 8 randomly chosen tutorial groups, was to give public recognition to students whose grade for the first midterm exam was within the top 30% of their group. We instructed the TAs in the control groups not to express any praise or criticism of the first midterm results in their groups.

We find that, compared to their peers in the control groups, the recipients of recognition in the treatment groups did no better. At the same time, the non-recipients who attended enough classes and were not too far off the cutoff grade for recognition significantly improved their performance, by up to 0.55 of the grade's standard deviation. Our findings taken together imply that unannounced recognition may be an effective motivational tool even when other powerful incentives are present, but its effectiveness depends on the characteristics of the target audience, such as class attendance and past performance, that it cannot influence.

In the rest of the paper, we review the existing literature (section 2), present our theoretical predictions (section 3), describe our experiment (section 4) and data (section 5) as well as relevant estimation issues. We report our empirical results in sections 6 and 7, and conclude with a summary of our findings and their implications in section 8.

2 Existing literature and our study

Our study builds on three related literatures – theories predicting recognition to affect performance, empirical research outside and within academia. Starting with the theories, recognition may affect performance by nurturing reciprocity between the agent and the principal (Akerlof, 1982), acting as a signal for the altruistic principal (Levine, 1998), providing information on the social norm to which people tend to conform (Bernheim, 1994) and activating status concerns (Moldovanu, Sela and Shi, 2007). However, our setting is not a principal-agent one: student effort does not directly affect teacher wealth, so teachers do not have a material incentive to stimulate it. This leaves two theoretical possibilities – conformity to the norm, and status concerns.

Under status concerns, recognition increases effort when receiving it leads to a higher status within a certain social group. Thus, Moldovanu et al.'s (2007) "contests for status" model predicts that in the presence of recognition every agent will put in effort proportionate to his or her ability rank within the group, whereas in the absence of recognition everyone's effort will be zero. Crucial for this mechanism to work is the expectation that recognition will occur in the future. Hence, effort response to an unannounced and one-off recognition could not be explained by status concerns.[1]

Conformity to the norm – a tendency to align actions to "a single standard of behavior despite heterogeneous preferences" (Bernheim, 1994, p. 841) – affects effort through coarse feedback on relative performance that comes in the form of recognition. Having received this feedback, the agents may adjust their beliefs about the norm and hence their effort choices. Specifically, the recipients of recognition will learn that they are more likely to have met the norm than they thought before, and will consequently reduce their effort, whereas the non-recipients will find themselves less likely to comply with the norm, and will therefore work harder. The opposite effort responses by the recipients and non-recipients of recognition is a marker of norm conformity being at work, distinguishing it from status concerns which encourage high performers to work progressively harder.

Turning to empirical research, providing recognition is found to affect a wide range of behaviors both within and outside the principal-agent setting: prosocial actions (Grant and Gino 2010), voluntary contributions (Chen, Harper, Konstan and Li, 2010), and output in short-term jobs (Kosfeld and Neckermann, 2011; Kube, Marechal and Puppe, 2012; Bradler, Dur, Neckermann and Non, 2016). One lesson from this research, which links it to the theory and is relevant for our experimental design, is that the effect of recognition depends on whether it is announced or spontaneous. Announced recognition affects behavior through status concerns, resulting in stronger responses from the already high performers (Kosfeld and Neckermann, 2011), whereas unannounced (or spontaneous) recognition, operating via conformity to the norm (Bradler et al., 2016), will have a larger effect on the relatively under-performing.

Lastly, there is research on recognition and feedback given to students. A large meta-

[1]While true in theory, this statement may not hold in some field settings, including ours, in which recognition may still be given in the future, albeit by a different person and for a different performance outcome. Then, once observed, the practice of recognition may increase the competitiveness of the environment, which in turn may trigger status concerns. We thank the anonymous referee for making this point.

study by Kluger and DeNisi (1996) concludes that, compared to other types of feedback (corrective feedback, progress assessments, reinforcement), "praise", the more frequently used term for recognition, is relatively ineffective. Hattie and Timperley (2007) mention the decrease in intrinsic motivation and dearth of learning-related information as the reasons for the lack of the effect. Based on their review of recognition literature, Marzano, Pickering and Pollock (2001) (pp. 53-58) conclude that recognition works best when it is given personally and for reaching a specified performance target. More recent studies on feedback generally find that it positively affects average student performance (Azmat and Iriberri, 2010; Tran and Zeckhauser, 2012; Bandiera, Larcinese and Rasul, 2015). They also find the effects of feedback to differ across the performance outcome distribution, being more pronounced for students who performed above and below average before receiving feedback.

The existing literature informs our theory and experimental design. Our theory rests on the idea of conformity to the norm, which predicts an effect of recognition through the provision of coarse feedback on relative performance. To try to isolate the influence of status concerns (even though it is hard to do so in the field), the recognition we give is unannounced. We give recognition based on a specified performance measure to all qualifying students, which is the best practice identified in the meta-studies in education psychology.

3 Theory and study hypotheses

As we want to investigate the effect of classroom recognition on subsequent performance, we apply the conformity to the norm theory – the only one applicable in our specific setting – to guide our empirical analysis and generate testable predictions. Assume agents in a group have a preference for conformity; that is, individual effort is positively affected by the performance norm in their group. The agents do not know precisely what the group norm is and base their effort choices on their beliefs about the norm. These beliefs are informed by signals each agent privately receives. Hence, all else equal, agents who initially received a low signal about the norm will work less hard than those who received a high signal. Feedback on relative performance in the form of recognition will correct the previously held beliefs about the norm and alter the effort choices as the result.

To formalize the norm beliefs update, we use the model from Bradler et al. (2016). In their model, the agent's optimal effort increases in the group norm γ. The effort directly and noiselessly transforms into output. The true γ is unobserved but it is common knowledge that it is drawn from a uniform distribution over $[0, 1]$. The agent chooses effort based on this knowledge and the privately received signal about γ.

There are two effort-choice periods in the model, one before and one after public recognition is given. In the pre-recognition period 1, the agent receives a signal $s_1 = \{High, Low\}$ about γ. The belief about the true γ is formed by this signal. In general,

$$E(\gamma|signal) = \frac{\int \gamma \cdot pr(signal|\gamma) f(\gamma) d\gamma}{\int pr(signal|\gamma) f(\gamma) d\gamma},$$

where $pr(signal|\gamma)$ is the conditional probability of receiving a certain signal, and $f(\gamma)$ is the probability density function of γ. Assuming that γ uniformly distributed over $[0, 1]$ and that $pr(s_1 = High|\gamma) = \gamma$ and $pr(s_1 = Low|\gamma) = 1 - \gamma$, the conditional expectations of γ given the signals are $E(\gamma|s_1 = High) = 2/3$ and $E(\gamma|s_1 = Low) = 1/3$. Thus, the agents who received the *High* signal will choose higher effort than those who received the *Low* signal.

In period 2, after recognition is publicly given to k out of n agents with the best performance results, agents update their beliefs about γ given the new signal $s_2 = \{Recognition,\ No\ Recognition\}$ each of them receives. Four combinations of signals received in two periods are possible: $(s_1 = High, s_2 = Recognition), (s_1 = High, s_2 = No\ Recognition), (s_1 = Low, s_2 = Recognition), (s_1 = Low, s_2 = No\ Recognition)$, each generating its own belief about γ.

Intuitively, the agents who put in high effort but did not receive recognition ($s_1 = High, s_2 = No\ Recognition$) will learn that there are at least k other agents who also put in high effort, so the group norm is probably high. The agents who put in low effort and did receive recognition nonetheless will learn that there are at least $n - k + 1$ other agents who also put in low effort. Recognition will lead these agents into believing that the norm is even lower than they thought, and subsequently reducing their effort. On the other hand, putting in low/high effort and not/receiving recognition is less surprising and changes the existing beliefs much less if at all.

The calculation of the conditional expectations of γ under different combinations of (s_1, s_2) involves aggregating statistical probabilities of s_2 and s_1 for an individual agent given the signals other agents might have received. Table 1 shows the expressions for conditional expectations of γ in periods 1 and 2 under the assumptions we have made regarding the probability of receiving the *High* or *Low* signals and the distribution of γ. It also reports the calculated values of conditional expectations for $n = 22$ and $k = 8$, our sample median group size and the number of recognition recipients. A detailed explanation with a simple example is provided in Appendix.

[Table 1 here.]

The results in Table 1 illustrate our intuitions about how the recipients and non-recipients of recognition will update their beliefs about the norm. The recipients will learn that the norm is probably less high than they initially believed, and will reduce their effort. On the other hand, the non-recipients will increase their effort, since their updated belief about the norm is higher than before. Hence, our first hypothesis:

H1: *Recognition will lead to an improvement in the performance of non-recipients, and to a deterioration in the performance of recipients.*

Also consistent with our intuition are the differences in the reactions of the recipients and non-recipients depending on whether they put in high or low effort in period 1, before recognition. The changes in the beliefs of the recipients who put in high effort and the non-recipients who put in low effort are small (0.59 vs. 2/3 for the high-effort recipients; 0.37 vs. 1/3 for the low-effort non-recipients), since their outcomes are highly predictable given their beliefs. Yet there are much bigger changes in the beliefs of the high-effort non-recipients (0.78 past-recognition vs. 2/3 pre-recognition) and the low-effort recipients (0.14 vs. 1/3), reflecting the surprise at their outcome in period 2.

The high-effort non-recipients's performance results are likely to be close to the cutoff starting from which recognition is given. For instance, in our adaptation of Bradler et al. (2016)'s model, half of the agents, on average, will receive the $High$ signal and choose high effort in period 1 but only about a third (=8/22) will receive recognition. The remaining 1/6 of the high-effort agents will not be recognized despite producing high output that would bring them close to the cutoff. This theoretical result informs our second hypothesis:

H2: *The non-recipients of recognition will show the greater improvement, the closer their pre-recognition performance is to the cutoff level starting from which recognition is given.*

One could also hypothesize that the recipients whose performance is just above the cutoff will reduce their effort by a larger margin than those high above the cutoff. The reason is that the recipients who put in low effort are closer to the cutoff. However, in our model with the expected 11 out of 22 agents putting in high effort and only 8 being recognized, it is unlikely that any recipient of recognition exerted low effort. With recognition being rather scarce in our study, we have neither sufficient statistical power to test this hypothesis nor a strong theoretical case for it.

4 Study context and experimental procedures

Our experiment involved first-year undergraduate students attending a compulsory microeconomics course at a large university in The Netherlands in the winter semester of 2012/13, which was the first semester for all students in our sample. It was taught in a way typical for a modern public university, which included regularly scheduled lectures given by the course leader to all students at once, as well as tutorials taught by teaching assistants (TAs) in tutorial groups. The tutorial groups of 20-30 randomly chosen students had been formed before the semester began, without the influence of any of the authors or the possibility to move between the groups. The tutorial groups were not the same for other courses.

There were two midterm exams weighing 10% each in the final grade, a weekly online test (10%), and the final exam at the end of the course (the remaining 70%). The first microeconomics midterm was chronologically the first midterm exam the students took. All elements of the course were graded according to the solutions key provided by the course lecturer. There was very little room for interpretation and no grading on the curve.

Our treatment took place between the two midterms and was given at randomly selected tutorial groups. There were ten TAs involved in the experiment, five of them teaching two groups each and the rest teaching one group. Before the start of the experiment, the TAs were invited to an information session where we explained the purpose of our experiment, requested their participation, and asked them not to tell about the experiment to the students. All TAs agreed to both requests. They were informed whether they were in a treatment or a control group after they had finished grading the first midterm. (The TAs who taught two groups had one treatment and one control group, randomly assigned.) The TAs in the treatment groups were asked to recognize in public the students whose first midterm grade was within the top 30% of their respective group's grade distribution. The TAs in the control groups were asked *not* to give any comments, positive or negative, on the first midterm results.

The procedure that the TA in the treatment groups were asked to adhere to is as follows. At the beginning of the first tutorial after the first midterm, the TA informed the students that their exam papers were to be handed out and discussed. Thereafter, he or she said the following, in Dutch: *I would like to have your attention for the students to whom I will now hand out their papers, as they did an excellent job. All of them received at least grade X* [the cutoff grade corresponding to the top 30% in the respective group, supplied by the authors]. *Experience tells us that students find the microeconomics midterms very hard. But these students did very well. My compliments!* After distributing those papers in random order, the TA continued by saying *I will now hand out the rest of the papers*, handing them out in random order as well. The tutorial then proceeded as usual, with each exam question being discussed and the exam papers collected in the end. There was no token or record that came along with the TA's message of recognition and therefore it could not be converted into future career or status benefits. The procedures in the control groups were exactly the same as in the treatment group (midterm papers were handed out in a random order, too), except that there were no recognition.

[Figure 1 here.]

Figure 1 gives a summary of our experimental procedures in chronological order. It is worth noting that the timing of the events is such that the last tutorial, when the students met their TAs for the last time, took place only 1 or 2 days after the second midterm exam. With this timing known to every course participant through official communications, there should be little expectation among students for the second midterm results to become available in time for recognition to recur. Therefore, while it cannot be completely excluded, the extent to which recognition can affect performance through status concerns is limited.

In sum, our study context is suitable for an experiment such as ours that looks at how the arrival of new information on relative performance changes beliefs about the norm. The information we provided in the form of recognition is new because our students received it in the very first semester of their studies, the microeconomics midterm being the first midterm exam they took. Also, since the tutorial groups varied by course, the students could not have learned it from participating in other courses in the same semester.

5 Data and estimation issues

Our study sample consists of 190 students in 8 treatment groups and 152 students in 7 control groups present on the (unannounced) treatment day. The 97 students not present on the treatment day differ from those in our sample. They attended fewer tutorials and were less likely to show up for the first midterm exam than those present. The 37 students absent on the treatment day who did write the first midterm received an average grade of 4.3 out of 10, significantly lower than our sample average of 5.8. Given these differences, we refrain from drawing conclusions for the student population in general, confining our analysis to finding the treatment effect on the treated.

There are 76% male and 24% female students in our sample, all aged between 18 and

20. The size of our sample was limited by the requirement of the course leader that all TAs should have at least one year of teaching experience. This requirement still leaves sufficient power for our statistical procedures to detect the effect of recognition if present. Thus, our power calculations for the overall effect of recognition based on its expected size (0.4 of standard deviation, as in Bradler et al. (2016)), the observed group average size (26) and dispersion (7.4), the intra-group correlation in grades (0.05), and the significance level of 0.05 show that we can detect this effect with probability 0.72. The same calculations for the effect on the non-recipients, assumed to be 0.5 standard deviation, produce the estimated power of 0.76.

The key outcome variable in our analysis is the second midterm exam grade, which in the Dutch education system ranges from 1 (the lowest) to 10 (the highest). We also examine other potentially interesting outcomes, obtained from administrative records, – the final exam grade, presence at tutorials, the probability of turning up for the midterm – in the extensions (section 6.2). [2]

We estimate the effect of our experimental treatment using the difference-in-difference estimator. In its basic version, it estimates the treatment effect as the average difference between post-treatment outcome change in the treatment and control groups:

$$\Delta y_i = \beta_0 + \beta_1 \cdot treatment_i + error_i, \tag{1}$$

where Δy_i is the outcome change for individual i and $treatment_i$ is the treatment dummy. Specification (1) can also estimate treatment effects separately for the recipients and non-recipients of recognition, as predicted by our Hypothesis 1, by interacting the treatment dummy with the one for recognition status. The would-be recipients (non-recipients) in the control groups are identified as falling within (outside) the top 30% of the first midterm grade distribution in their group.

The two estimation issues applicable to equation (1) are i) the possible bias to the estimate of the treatment effect (β_1) due to regression to the mean, and ii) the cluster structure of the error term, which invalidates the conventional OLS estimates of the treatment effect's standard error. Regression to the mean is likely to occur when, despite randomization, the

[2] In addition to the above data, we tried to obtain information on the hours spent studying for the course at home with questionnaires distributed before and after the treatment on behalf of the author not involved in the course. However, the response rate on the study hours question was less than 30% and, for those who did respond, the correlation with their grades was about zero, casting doubt on the reliability of the answers.

control and treatment groups have different averages of the baseline outcome. In this case, the natural convergence of the outcomes in the control and treatment groups to a common mean will be mistaken for the treatment effect (Stigler, 1997). To see this, note that (1) is a restricted version of the more general equation

$$\Delta y_i = \beta_0 + \gamma \cdot y_{i,-1} + \beta_1 \cdot treatment_i + error_i, \tag{2}$$

where $y_{i,-1}$ is the pre-treatment outcome and γ is a parameter that measures the speed of convergence to the common mean $\frac{-\beta_0}{\gamma}$. The OLS estimate of the treatment effect from (1),

$$\widehat{\beta_1} = \beta_1 + \gamma \frac{\mathrm{cov}(y_{i,-1}, treatment_i)}{\mathrm{var}(treatment_i)},$$

will be biased unless $\gamma = 0$ (no convergence) or the average pre-treatment grades in the treatment and control groups are equal. As we show in the next section (Tables 2 and 3), neither of these conditions is satisfied in our data. As a solution to remove the bias to the treatment effect's estimate, we run the generalized difference-in-difference equation (2) in addition to the baseline equation (1).

The second estimation issue – clustered standard errors – applies because our treatment is given at the tutorial group level. If there is a correlation between regression errors within tutorial groups, say via group-specific unobservables such as TA quality, OLS will underestimate the treatment effect standard error by a factor proportional to the strength of the within-group correlation in errors and to the number of observations in a group (Moulton, 1986). Angrist and Pischke (2009, chapter 8) discuss several approaches to repairing the regression coefficients' standard errors in the presence of clustering, which range from allowing for inter-cluster correlation in the residual variance-covariance matrix (Liang and Zeger, 1986; Stata `cluster` option) to running (weighted) regressions in group averages (Donald and Lang, 2007), to bootstrapping at the group level, or "wild bootstrap" (Cameron, Gelbach and Miller, 2008). We have implemented all these methods and found bootstrapping to give the most conservative p-values, which we report in the regression tables.

6 Baseline results

6.1 The average treatment effect on the midterm exam grade

One gets the first impression of the treatment effect from Figure 2 that plots cumulative distributions for the second to first midterm grade change for the entire sample as well as for the recipients and non-recipients of recognition. There is no difference between the control and treatment groups around the median grade change of 0. Yet, there is a shift in the mass of the distribution to the right, which is most noticeable for the non-recipients of recognition with visibly fewer negative grade changes. In fact, the Wilcoxon rank-sum test rejects the null hypothesis of the equality of the treatment and control grade change distributions with a p-value of 0.08 for all and 0.06 for the non-recipients.

[Figure 2 here.]

To quantify the treatment effect, we first turn to Table 2 which reports the averages and standard deviations of the midterm grade by period (before or after the treatment), group (control or treatment), and status (all, recipients and non-recipients), and then to Table 3 which reports regression estimates for the treatment effect from equations (1) and (2). The average grades are fairly low, being just above the pass level (5.5) and below the pass for the non-recipients. There is a slight deterioration in the control group grades from the first to second midterm (6.30 to 6.11), due to recipients' average grade going down, and an improvement in the treatment group grades (5.63 to 5.94) thanks to improved performance of the non-recipients.

[Table 2 here.]

The simple difference-in-difference estimates of effect of recognition on midterm grade change is about 0.5 for the entire sample and 0.6 for non-recipients (Table 3). Both estimates are fairly large relative to the grade's standard deviation (about 2.3). However, accounting for the cluster structure of the error term in equation (1) by bootstrapping, we see that the treatment effect estimates are not precise enough to be considered statistically significant at conventional levels.

[Table 3 here.]

We also observe from Table 2 that there are considerable pre-treatment differences between the control and treatment groups, especially for the non-recipients. These differences go down post-treatment, suggesting that part of the estimated treatment effect may in fact be due to regression to the mean. To account for this possibility, we run the generalized difference-in-difference estimator (equation 2). The results show that the parameter γ, omitted in equation (1), is in fact large (about -0.4) and significant, implying that there is indeed convergence between the pre-treatment grade distributions in the control and treatment groups which would have happened even in the absence of treatment. Taking this convergence into account, the treatment effect estimates are smaller than those previously reported: 0.24 for all, 0.08 for recipients, and 0.35 for non-recipients. None is statistically significant. The difference between the treatment effect estimates for the non-recipients and recipients, -0.26, is insignificant, too (bootstrap p-value=0.58), lending little support to our Hypothesis 1.

Remark: Of the 359 students who attended the first midterm and were present when recognition was given, 17 did not show up for the second midterm (7 in the treatment, 10 in the control group) and were therefore not included in our analysis, most of which uses the grades from both midterms (hence 359-17=342 observations in our sample). Could this sample attrition have influenced our results? While the students who do not show up for the second midterm have a significantly lower first midterm grade than those who do (3.9 vs. 5.4), there is no treatment effect on attrition, either for recipients or non-recipients. Hence the positive effect of sample attrition on the average grade is balanced across the treatment groups and is therefore cancelled out in our treatment effect estimates.

6.2 The average treatment effect on other outcomes

We now analyze the effect of recognition on the final exam grade, presence at the tutorials, and the probability of turning up for the second midterm. All these outcomes are related to student effort and can therefore be affected. Table 4 reports the estimation results from the generalized difference-in-difference equation (2) estimated for each of those variables. [3]

[Table 4 here.]

There seems to be an effect on the recipients' probability of turning up for the second midterm, which increases by 5 percentage points (ppts) after the treatment. However, with everyone in the control group turning up for both midterms, the treatment effect's standard error is downward-biased. Besides, whether it is significant or not, relative to the pre-treatment exam attendance of more than 90% the effect of 5 ppts is very small. Note that the treatment effect on the probability of turning up is not the same as that on

[3] Unlike presence at tutorials and midterms, the final exam grade is not available before treatment. We estimate equation (2) with final exam grade on the left hand-side and the treatment dummy and the first midterm grade on the right.

sample attrition, since the underlying sample now includes students present as well as not present at the first midterm.

There is no treatment effect on the final exam grade, perhaps because of the uniformly high importance of the final exam, whose weight is 70% of the total grade, for all students regardless of recognition status. Indeed, there is a lot less variation in the final exam grade (std. deviation 1.7) than in the midterm grades (std. deviation about 2.4). The effect of our rather delicate treatment on the final grade may simply be overwhelmed by more important, and uniform, concerns.

Importantly for our further analysis, presence at tutorials is unaffected by the treatment, suggesting that tutorial attendance is the outcome influenced by student preferences rather than behavior toward conformity. To the extent that tutorial attendance is informative of the degree of importance students attach to the course, we can use it to see how it influences the effect of recognition on student performance. Its exogeneity with respect to the treatment allows us to incorporate this variable in our analysis in a simple way as we do in the next section.

7 Treatment effect heterogeneity

Recall Figure 2 that shows the treatment/control differences in the distributions of the non-recipients' second to midterm grade change to be more pronounced off the median, suggesting that the treatment effect is local to some parts of the grade distribution. Indeed, running a bootstrapped quantile regression estimator on specification (2) for the non-recipients gives the treatment effect estimates (p-values) of 0.625 (0.290), 0.071 (0.810) and 1 (0.038) for the first quartile, median and the third quartile of the non-recipients' grade distribution. The treatment effect being stronger for the non-recipients closer to the recognition cutoff grade is consistent with our theory which predicts a stronger treatment effect for the non-recipients closer to the recognition cutoff (Hypothesis 2). In this section we further explore treatment effect heterogeneity as predicted by our theory. Additionally, we test whether the effect of recognition varies with the interest in the course, which we proxy with tutorial attendance itself found to be unaffected by recognition. Unless stated otherwise, the results in this section are exclusively for the non-recipients of recognition.

Starting with treatment effect heterogeneity by the distance to the recognition cutoff (d), one approach to detecting it is to interact the treatment dummy in equation (2) with d. Doing so gives a treatment effect estimate of 1.06 (p-value$=0.15$) and the coefficient on the $treatment * d$ interaction term of -0.24 (p-value$=0.2$). Consistent with Hypothesis

2, these results, though weak, suggest a reduction in the power of recognition with the distance to the cutoff.

The reduction in the treatment effect with the distance to cutoff does not have to be linear; in fact, the non-recipients may not react at all or even withdraw effort when the distance to the cutoff is past some threshold. Because theory offers little guidance in establishing this threshold, we rely on the data for clues. We choose 1.5 grade points as the threshold distance, which is the median of the distribution of the first to second midterm grade improvements of the would-be non-recipients in the control group. We estimate the treatment effects for the non-recipients whose distance to the recognition cutoff is above and below this threshold distance. We also re-estimate the treatment effects with the threshold distances of 1 and 2 grade points as a robustness check.

Table 5 reports the effects of recognition on the non-recipients with the distance to the cutoff grade below and above the threshold. The treatment effect for the non-recipients with $d \leq 1.5$ is 0.91 (p-value=0.05), whereas the effect for those with $d > 1.5$ is much smaller, 0.05 (p-value=0.91). Repeating the same exercise for the alternative threshold values gives the estimated effects of 0.53 (p-value=0.5) and 0.29 (p-value=0.5) for the non-recipients with $d \leq 1$ and $d > 1$, and 0.59 (p-value=0.2) and 0.15 (p-value= 0.75) for those with $d \leq 2$ and $d > 2$, respectively.

[Table 5 here.]

We now turn to the interest in the course as another moderator of the effect of recognition. Although everyone is under pressure to pass the course, the importance attached to it is likely to vary by student. Though not directly observed to us, it can be inferred from tutorial attendance with the help of a simple theoretical argument. In theory, the frequency of tutorial attendance is an outcome of each student's choosing a utility-maximizing tradeoff between the amount of time spent studying microeconomics and other, competing, activities. Hence, more tutorials attended indicates higher interest in the course.

In practice, this tradeoff may be affected by the requirement to attend at least 10 out of the total of 13 tutorials in order to be admitted to the final exam and hence to complete the course. This requirement is reflected in the fact that only 2% of the students in our sample attended fewer than 10 tutorials compared to 14% who attended exactly 10 tutorials. However, a further 20%, 27% and 36% attending 11, 12 and 13 tutorials respectively, thus exceeding the attendance requirement voluntarily, suggests that the tutorial attendance requirement is not binding for the majority of students.

[Table 6 here.]

Table 6 reports the treatment effects by tutorial attendance. We choose the threshold tutorial attendance for the non-recipients at 12 out of max. 13. About 50% of students, both recipients as well as non-recipients, attended 12 or more tutorials. The treatment effect for those attending 12 or more tutorials is 0.69 (p-value=0.05), and -0.28 (p-value= 0.56) for those attending fewer than 12. The treatment effects with the threshold of 11 are similar – 0.65 (p-value= 0.08) at or above the threshold, and -1.02 ($p-$value= 0.21) below – and show that the effect of recognition becomes fairly large and statistically significant once we exclude only 41 unmotivated students from the sample. [4]

We now incorporate both sources of heterogeneity into a single regression specification, which we also estimate with additional controls for group size, mean first midterm grade, and its variance. The results, reported in Table 7, show that the treatment effect for the non-recipients increases with both tutorial attendance and proximity to the cutoff grade, reaching 1.38 (p-value=0.02) for those who attended 12 or more tutorials and were within $d \leq 1.5$ distance from the cutoff grade for recognition. They make up 23% of all non-recipients. Outside this group, the treatment effect is insignificant. Controlling for group characteristics, the most important of which turns out to be group size, produces qualitatively similar results: a treatment effect of 1.29 (p-value=0.01) for the non-recipients who attended 12 or more tutorials and stayed within $d \leq 1.5$, and a much smaller and less significant effect for the rest of the sample. Though local to a specific group of students in our sample, the treatment effect of 1.29 is substantial and is equivalent to 0.55 of the first midterm grade's standard deviation.

[Table 7 here.]

Applying the empirical analysis in this section to the outcomes other than second midterm grade (final exam grade, post-recognition tutorial attendance, probability of showing up at the second midterm exam – section 6.2) brings no significant results. Hence, there is neither average nor local treatment effects of recognition on these outcomes.

8 Discussion and conclusion

We set out to investigate the effect of classroom recognition on subsequent performance. There are several theories that predict such an effect, which we discuss in section 2. Therein

[4] We have also tried the tutorial attendance only prior to the treatment as an alternative measure of the interest in the course. Repeating the same analysis as above, we found no significant heterogeneity in the effect of recognition. The variation in pre-recognition tutorial attendance is very low with 85% of students attending 7 or 8 tutorials, which casts doubt on the ability of this measure to capture variation in the interest in the course.

we also state that for a non-principal-agent setting such as ours there are only two relevant theories: status concerns and conformity to the norm. The status concerns theory does not apply in our setting because the recognition we give is unexpected and strictly one-off. This leaves only one theory – conformity to the norm – which we use to generate predictions (section 3).

Testing these predictions with data generated through our field experiment has produced three main results: 1) there is no effect of recognition on its recipients' further performance; 2) there is a locally positive effect on the non-recipients' performance; and 3) the effect for the non-recipients increases with tutorial attendance and the proximity to the cutoff grade for recognition. The first two results are consistent with those in Bradler et al. (2016), the field experiment closest to ours, but in a setting quite different from theirs, since it involves people doing their main activity and having powerful incentives other than recognition. Our third finding – that distance to the cutoff for recognition and interest in the task moderate the performance effect of recognition – is new. It implies that recognition improves student performance only when there is a big enough update in beliefs about the norm and a high enough interest in the activity being recognized. Under these conditions the effect of recognition is quantitatively large, about half of the grade's standard deviation, but tails off to insignificance for a large part of our sample.

Our findings are consistent with the prediction of the conformity theory that the effect of recognition on non-recipients' effort is stronger than than on the recipients' (at least for the majority of the sufficiently motivated students, recall Table 6). This theory is applicable to our experimental setting because we work with unannounced and spontaneous recognition. Yet, some of our empirical results cannot be explained by conformity alone. For instance, although conformity predicts a negative response from the recipients (Hypothesis 1), we do not observe this, even when we control for tutorial attendance, our measure of interest in the course. It seems, therefore, that our findings are the product of a mix of factors, conformity to the norm being only of them. Another factor could be intrinsic motivation, since there was no deterioration in the recipients' performance. Yet another is likely to be institutional pressures to perform – hence the lower variation in the final exam grade and no treatment effect on it. A rigorous test of the conformity theory will have to control for these factors.

We believe our results have important implications for the research in, and practice of, education. The finding that it is non-recipients who respond to our treatment implies that spontaneous recognition of the sort we have implemented may be just the right tool to stimulate the performance of currently underachieving students, as long as they are not too far behind. It may thus be preferable to more formal types of recognition based on tournaments (for example, dean's lists and other forms of academic distinction) which stimulate to a greater extent the performance of the already more able students (Kosfeld and Neckermann, 2011). However, care needs to be taken in deciding on the cutoff grade for recognition, since, as Bradler et al. (2016) showed, too low a cutoff reduces the information content in recognition, and hence the response to it, while too high a cutoff makes recognition unattainable as well as uninformative.

We conclude by reflecting on the limitations of our study, and by stating some outstanding questions that await further research. Starting with limitations, the biggest one is that our experiment has turned out to be under-powered. In our power calculations, done prior to administering the treatment, we assumed a treatment effect commensurate to that found in Bradler et al. (2016), in which case our experiment would have had sufficient power. The actual treatment effect, however, is much smaller. More experiments like ours are needed to obtain clearer results. Fortunately, they are not very difficult to set up when one has regular access to a large student audience.

As for the outstanding questions, the realism of our experiment weakens our ability to fully capture the mechanisms behind our findings, as well as to control for potentially important unobservables, which are all the more important given the heterogeneity in the treatment effect. Does the importance of the group norm vary by group? What is the role of the identity of the recognition giver? Will the effect of recognition disappear when students learn their relative performance within the group? Answering these questions through further research will help better understand the workings of recognition and may be useful in maximizing its benign effects.

9 Appendix

Here we explain the calculations of beliefs before and after recognition (table 1) with a simple example. Suppose there are $n = 4$ agents in a group and $k = 2$ receive recognition. For the agent with ($s_1 = High, s_2 = Recognition$),

$$E\left(\gamma | s_1 = High, s_2 = Recognition\right) = \frac{\int_0^1 \gamma \cdot pr(s_1 = High \,\&\, s_2 = Recognition) d\gamma}{\int_0^1 pr(s_1 = High \,\&\, s_2 = Recognition) d\gamma}$$

$$= \frac{\int_0^1 \gamma \cdot \left(\gamma(1-\gamma)^3 + \binom{3}{1}\gamma^2(1-\gamma)^2 + \binom{3}{2}\frac{2}{3}\gamma^3(1-\gamma) + \frac{2}{4}\gamma^4\right) d\gamma}{\int_0^1 \left(\gamma(1-\gamma)^3 + \binom{3}{1}\gamma^2(1-\gamma)^2 + \binom{3}{2}\frac{2}{3}\gamma^3(1-\gamma) + \frac{2}{4}\gamma^4\right) d\gamma} = \frac{\frac{13}{60}}{\frac{7}{20}} = \frac{13}{21}$$

The expression $pr(s_1 = High \,\&\, s_2 = Recognition) = \gamma(1-\gamma)^3 + \binom{3}{1}\gamma^2(1-\gamma)^2 + \binom{3}{2}\frac{2}{3}\gamma^3(1-\gamma) + \frac{2}{4}\gamma^4$ is the sum of the probabilities of all the events when an agent who received $s_1 = High$ and exerted high effort will receive recognition afterwards. Its first term, $\gamma(1-\gamma)^3$, is the probability that only one person – the agent himself – received $s_1 = High$. The second term is the probability that one more agent received $s_1 = High$. It contains the coefficient $\binom{3}{1} = 3$ to include all three possibilities. As long as there are one or two out of four agents who received $s_1 = High$, they both will get recognition. The third and fourth terms are the probabilities that two or three other agents also received $s_1 = High$. When two or three other agents receive $s_1 = High$, the agent's chances to receive recognition are $\frac{2}{3}$ and $\frac{2}{4}$, respectively, since only two agents will be given recognition. Hence the coefficients on the last two terms.

For the agent with ($s_1 = High, No\ Recognition$),

$$E\left(\gamma | s_1 = High, s_2 = No\ Recognition\right) = \frac{\int_0^1 \gamma \cdot pr(s_1 = High \,\&\, s_2 = No\ Recognition) d\gamma}{\int_0^1 pr(s_1 = High \,\&\, s_2 = No\ Recognition) d\gamma}$$

$$= \frac{\int_0^1 \gamma \cdot \left(\binom{3}{2}\frac{1}{3}\gamma^3(1-\gamma) + \frac{2}{4}\gamma^4\right) d\gamma}{\int_0^1 \left(\binom{3}{2}\frac{1}{3}\gamma^3(1-\gamma) + \frac{2}{4}\gamma^4\right) d\gamma} = \frac{\frac{7}{60}}{\frac{3}{20}} = \frac{7}{9}$$

Notice that the agent with $s_1 = High$ may not receive recognition only if there are at least two more agents also with $s_1 = High$. In the expression $pr(s_1 = High \,\&\, s_2 = No\ Recognition) = \binom{3}{2}\frac{1}{3}\gamma^3(1-\gamma) + \frac{2}{4}\gamma^4$, the first and second terms are the probabilities that two and three more agents received $s_1 = High$, times the probabilities that the agent in question did not get recognition.

For the agent with $(s_1 = Low, s_2 = Recognition)$,

$$E\left(\gamma | s_1 = Low, s_2 = Recognition\right) = \frac{\int_0^1 \gamma \cdot pr(s_1 = Low \,\&\, s_2 = Recognition) d\gamma}{\int_0^1 pr(s_1 = Low \,\&\, s_2 = Recognition) d\gamma}$$

$$= \frac{\int_0^1 \gamma \cdot \left(\frac{2}{4}(1-\gamma)^4 + \binom{3}{1}\frac{1}{3}\gamma(1-\gamma)^3\right) d\gamma}{\int_0^1 \left(\frac{2}{4}(1-\gamma)^4 + \binom{3}{1}\frac{1}{3}\gamma(1-\gamma)^3\right) d\gamma} = \frac{\frac{1}{30}}{\frac{3}{20}} = \frac{2}{9}$$

The agent who received $s_1 = Low$ can receive recognition in two cases. First, when all other agents also received $s_1 = Low$, in which case two randomly picked agents will receive recognition (hence the chances of receiving recognition are $2/4$). Second, when only one other agent receives $s_1 = High$, in which case the agent can get the remaining recognition place with probability $1/3$. The first and second terms in the expression $pr(s_1 = Low \,\&\, s_2 = Recognition)$ correspond to these cases.

Finally, for the agents who received $s_1 = Low$ and did not get recognition,

$$E\left(\gamma | s_1 = Low, s_2 = No\ Recognition\right) = \frac{\int_0^1 \gamma \cdot pr(s_1 = Low \,\&\, s_2 = No\ Recognition) d\gamma}{\int_0^1 pr(s_1 = Low \,\&\, s_2 = No\ Recognition) d\gamma}$$

$$= \frac{\int_0^1 \gamma \cdot \left(\gamma^3(1-\gamma) + \binom{3}{1}\gamma^2(1-\gamma)^2 + \binom{3}{2}\frac{2}{3}\gamma(1-\gamma)^3 + \frac{2}{4}(1-\gamma)^4\right) d\gamma}{\int_0^1 \left(\gamma^3(1-\gamma) + \binom{3}{1}\gamma^2(1-\gamma)^2 + \binom{3}{2}\frac{2}{3}\gamma(1-\gamma)^3 + \frac{2}{4}(1-\gamma)^4\right) d\gamma} = \frac{\frac{2}{15}}{\frac{7}{20}} = \frac{8}{21}$$

In the expression $pr(s_1 = Low \,\&\, s_2 = No\ Recognition) = \gamma^3(1-\gamma) + \binom{3}{1}\gamma^2(1-\gamma)^2 + \binom{3}{2}\frac{2}{3}\gamma(1-\gamma)^3 + \frac{2}{4}(1-\gamma)^4$, the first term is the probability that all three other agents received $s_1 = High$, the second term is the probability that one more agent received $s_1 = Low$. In both cases, the agent in question would certainly not receive recognition. The third term is the probability that two more agents received $s_1 = Low$, in which case the probability of not receiving recognition is $2/3$. The fourth term is the probability that all agents received $s_1 = Low$, in which case the chances of receiving and not receiving recognition are equal.

References

Akerlof, G.A. 1982. "Labor contracts as a partial gift exchange." *Quarterly Journal of Economics*, 97(4): 543–569.

Angrist, J., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Azmat, A., and N. Iriberri. 2010. "The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment Using High School Students." *Journal of Public Economics*, 94: 435–452.

Bandiera, O., V. Larcinese, and I. Rasul. 2015. "Blissful ignorance? Evidence from a natural experiment on the effect of individual feedback on performance." *Labour Economics*, 34: 13–25.

Bernheim, D. 1994. "A Theory of Conformity." *Journal of Political Economy*, 102: 841–877.

Bradler, S., R. Dur, S. Neckermann, and A. Non. 2016. "Employee Recognition and Performance: A Field Experiment." *Management Science*, forthcoming.

Cameron, A., J. Gelbach, and D. Miller. 2008. "Bootstrap-based improvements for inference with clustered errors." *The Review of Economics and Statistics*, 90: 414–427.

Chen, Y., F. Harper, J. Konstan, and S. Li. 2010. "Social comparisons and contributions to online communities: A field experiment on movielens." *American Economic Review*, 100: 1358–1398.

Donald, S., and K. Lang. 2007. "Inference with difference-in-difference and other panel data." *The Review of Economics and Statistics*, 89: 221–233.

Grant, A., and F. Gino. 2010. "A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior." *Journal of Personality and Social Psychology*, 98: 946–955.

Harrison, G., and J. List. 2004. "Field experiments." *Journal of Economic Literature*, 42: 1009–1055.

Hattie, J., and H. Timperley. 2007. "The power of feedback." *Review of Educational Research*, 77: 81–112.

Kluger, A.N., and A. DeNisi. 1996. "The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory." *Psychological Bulletin*, 119: 254–284.

Kosfeld, M., and S. Neckermann. 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics*, 3: 86–99.

Kube, S., M. Marechal, and C. Puppe. 2012. "The Currency of Reciprocity: Gift Exchange in the Workplace." *American Economic Review*, 102: 1644–1662.

Levine, D. 1998. "Modelling altruism and spitefullness in experiments." *Review of Economic Dynamics*, 1: 593–622.

Levitt, S., J. List, S. Neckermann, and S. Sadoff. 2016. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." *American Economic Journal: Economic Policy.*

Liang, K.-Y., and S. Zeger. 1986. "Longitudinal data analysis using generalsed linear models." *Biometrica*, 73: 13–22.

Marzano, R., D. Pickering, and J. Pollock. 2001. *Classroom Instruction that Works: Research-based Strategies for Increasing Student Achievement.*

Moldovanu, B., A. Sela, and X. Shi. 2007. "Contests for Status." *Journal of Political Economy*, 115: 338–363.

Moulton, B. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics*, 32: 385–397.

Stigler, S. 1997. "Regression towards the mean, historically considered." *Statistical Methods in Medical Research*, 6: 103–114.

Tran, A., and R. Zeckhauser. 2012. "Rank as an inherent incentive: Evidence from a field experiment." *Journal of Public Economics*, 96: 645–650.

Figure 1: Timeline of the experiment (in weeks)

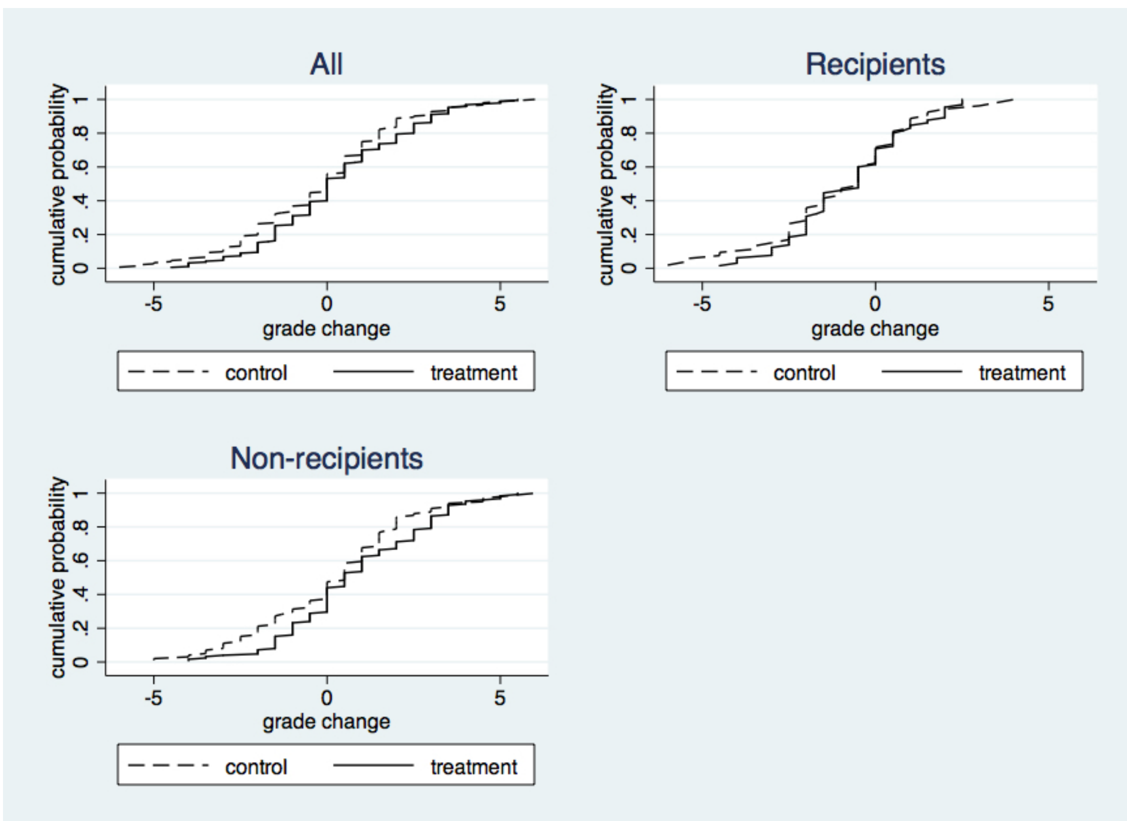

Figure 2: Second to first midterm grade change distribution

Table 1: Beliefs about the norm before and after recognition

Round 1: before recognition			Round 2: after recognition		
Signal s_1	Belief $E\left(\gamma \mid s_1\right)$	Value	Signal s_2	Belief $E\left(\gamma \mid s_1, s_2\right)$	Value
High	$\frac{\int_0^1 \gamma^2 d\gamma}{\int_0^1 \gamma d\gamma}$	2/3	*Recognition*	$\frac{\int_0^1 \gamma \cdot \left(\sum_{i=1}^{n} \gamma^i (1-\gamma)^{n-i} \binom{n-1}{i-1} \min(1,k/i)\right) d\gamma}{\int_0^1 \left(\sum_{i=1}^{n} \gamma^i (1-\gamma)^{n-i} \binom{n-1}{i-1} \min(1,k/i)\right) d\gamma}$	0.59
			No Recognition	$\frac{\int_0^1 \gamma \cdot \left(\sum_{i=k+1}^{n} \gamma^i (1-\gamma)^{n-i} \binom{n-1}{i-1} \frac{i-k}{i}\right) d\gamma}{\int_0^1 \left(\sum_{i=k+1}^{n} \gamma^i (1-\gamma)^{n-i} \binom{n-1}{i-1} \frac{i-k}{i}\right) d\gamma}$	0.78
Low	$\frac{\int_0^1 \gamma(1-\gamma) d\gamma}{\int_0^1 (1-\gamma) d\gamma}$	1/3	*Recognition*	$\frac{\int_0^1 \gamma \cdot \left(\sum_{i=0}^{k-1} \gamma^i (1-\gamma)^{n-i} \binom{n-1}{i} \frac{k-i}{n-i}\right) d\gamma}{\int_0^1 \left(\sum_{i=0}^{k-1} \gamma^i (1-\gamma)^{n-i} \binom{n-1}{i} \frac{k-i}{n-i}\right) d\gamma}$	0.14
			No Recognition	$\frac{\int_0^1 \gamma \cdot \left(\sum_{i=0}^{n-1} \gamma^i (1-\gamma)^{n-i} \binom{n-1}{i} \min\left(1, \frac{n-k}{n-i+1}\right)\right) d\gamma}{\int_0^1 \left(\sum_{i=0}^{n-1} \gamma^i (1-\gamma)^{n-i} \binom{n-1}{i} \min\left(1, \frac{n-k}{n-i+1}\right)\right) d\gamma}$	0.37

Note: The beliefs are calculated for the group size $n = 22$ and the number of recognition recipients $k = 8$.

Table 2: Descriptive statistics

	All		Recipients		Non-Recipients	
	Control	Treatment	Control	Treatment	Control	Treatment
Grade before	6.30	5.63	8.44	8.09	5.15	4.35
	(2.38)	(2.25)	(1.40)	(1.04)	(1.97)	(1.52)
Grade after	6.11	5.94	7.50	7.36	5.37	5.20
	(2.49)	(2.47)	(2.13)	(1.95)	(2.36)	(2.39)
No. of observations	152	190	53	65	99	125

Notes: Recipients (non-recipients) in the control groups are defined as the students who *would be* recognized (not recognized), that is, those within (outside) the top 30% of their group's first midterm grade distribution. Standard deviations are in parentheses.

Table 3: The effect of recognition on student performance: baseline regression results

	All		Recipients		Non-Recipients	
	Eq. (1)	Eq. (2)	Eq. (1)	Eq. (2)	Eq. (1)	Eq. (2)
Treatment effect (β_1)	0.49	0.24	0.21	0.08	0.62	0.35
	(0.35)	(0.32)	(0.44)	(0.39)	(0.40)	(0.40)
Convergence parameter (γ)		-0.37***		-0.39**		-0.34***
		(0.04)		(0.07)		(0.08)
No. of observations	342		118		224	

Notes: Dependent variable: Change in the midterm grade. Block bootstrap standard errors calculated using Cameron, Gelbach and Miller (2008)'s method based on 1000 repetitions are in parentheses. * , **, *** denote coefficients significant at 1%, 5% and 10% significance levels.

Table 4: Estimated treatment effects for other outcomes that could be affected by recognition

	All	Recipients	Non-Recipients
		Final exam grade	
Treatment effect (β_1)	0.02	0.19	-0.04
	(0.24)	(0.20)	(0.33)
No. of observations	336	117	219
		Presence at tutorials	
Treatment effect (β_1)	0.06	0.16	0.02
	(0.12)	(0.19)	(0.13)
No. of observations	342	118	224
		Probability of turning up for the second midterm	
Treatment effect (β_1)	0.03	0.05**	0.02
	(0.025)	(0.023)	(0.03)
No. of observations	368	121	247

Notes: All estimates are based on equation (2) controlling for regression to the mean. Bootstrap standard errors are in parentheses. * , **, *** denote coefficients significant at 1%, 5% and 10% significance levels.

Table 5: Treatment effects by the distance to cutoff grade (d), non-recipients only

	$0 < d \leq 1.5$	$d > 1.5$
Treatment effect	0.91**	0.05
	(0.47)	(0.46)
No. of observations	77	147
	$0 < d \leq 1$	$d > 1$
Treatment effect	0.53	0.29
	(0.78)	(0.45)
No. of observations	55	169
	$0 < d \leq 2$	$d > 2$
Treatment effect	0.59	0.15
	(0.51)	(0.46)
No. of observations	108	116

Notes: Dependent variable: second midterm grade. All estimates are based on equation (2) controlling for regression to the mean. Bootstrap standard errors based on 1000 repetitions are in parentheses. * , **, *** denote coefficients significant at 1%, 5% and 10% significance levels.

Table 6: Treatment effects by tutorial attendance, non-recipients

	<12 tutorials	12 or more tutorials
Treatment effect	-0.28	0.69**
	(0.48)	(0.35)
No. of observations	93	131

	<11 tutorials	11 or more tutorials
Treatment effect	-1.02	0.65*
	(0.77)	(0.38)
No. of observations	41	183

Notes: Dependent variable: second midterm grade. All estimates are based on equation (2) controlling for regression to the mean. Bootstrap standard errors based on 1000 repetitions are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Treatment effects by tutorial attendance and the distance to cutoff grade (d), non-recipients

	Non-recipients	
	$0 < d \leq 1.5$	$1.5 < d$
< 12 tutorials attended	-0.14 (0.65)	-0.25 (0.55)
No. obs. in group	26	67
12 or more tutorials attended	1.38** (0.59)	0.16 (0.41)
No. obs. in group	51	80
The same controlling for group size, mean & st.dev. grade	1.29*** (0.50)	-0.09 (0.34)
Log group size	1.92** (0.79)	
Group average grade	0.16 (0.28)	
Grade std. deviation	-0.67 (0.51)	

Notes: Dependent variable: second midterm grade. All estimates are based on equation (2) controlling for regression to the mean. Bootstrap standard errors based on 1000 repetitions are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.