Pre-analysis plan of the project:

The effects of incentivized goals on academic performance

Pol Campos-Mercade[*] and Erik Wengström[*]

February 11th, 2018

Abstract

We incentivize university students with €300 conditional on attaining a goal GPA (Grade Point Average). First, we will study both the average treatment effect and the (potentially different) treatment effects throughout the distribution. Second, we will study the peer-effects of the policy. Third, we will complement these results with data from an incentivized survey, in which we measure students' WTP (Willingness to Pay) for this policy, GPA beliefs, and preferences. These data allow us to further understand the mechanisms of the treatment effect and to perform welfare analyses.

Keywords: Incentives, pay for performance, goals, peer-effects.

JEL classification: C93, I22.

[*] Lund University, Department of Economics, Tycho Brahes väg 1, Lund, Sweden. E-mails: pol.campos@nek.lu.se and erik.wengstrom@nek.lu.se.

This document was uploaded on the AEA RCT Registry on February 12th, 2018, right before the beginning of the data collection. We chose to upload it as a document, instead of written directly on the website, to prevent the study participants from learning about the design and purpose of this experiment.

Primary Outcome

Second semester GPA: average grade of the five second semester courses (Microeconomics, Mathematics II, Business Administration, Accountability and Economic History).

Secondary Outcomes to study the treatment effects on academic performance

Secondary variables used as outcomes to further study the treatment effects on students' academic performance:

- *Number of courses passed in the second semester.*
- *Third and fourth semester GPA*: average grade of their ten second year courses.
- *Number of courses passed in the third and fourth semester.*
- *Total number of courses passed during their first two years of studies.*
- *Average grade in Mathematics II and Microeconomics.*[1]
- *Grades for each of the different second semester courses.*
- *Average grades of the multiple-choice parts of the exams.*[2]

Secondary Outcomes to study how students perceive the treatment

Variables used as outcomes to study heterogeneity in how students perceive the incentivized goals:

- *Willingness to Pay (WTP)*: The questionnaire contains incentivized questions aimed at capturing the participant's WTP to receive the incentivized goal
- *The goal's non-monetary value*: WTP minus the expected monetary value of the goal. Our preferred way to calculate the expected monetary value of the goal is to use a question in the questionnaire in which we ask whether they prefer €300 with p probability (and €0 with $1 - p$ probability) or some amount of money for sure. Here, p is the participants' own estimate of the probability that they will obtain a GPA of 6 or higher if they are treated. A second way is to use the expected monetary value of the goal, €300*p. Here we arguably leave open some degrees of freedom to decide. We will choose the first option if the measure is not too noisy. Otherwise, we will choose the second one.
- *Believed personal treatment effect*: defined as the difference between the GPA that a student believes she will get with the goal minus the believed GPA without the goal. We will use the belief elicitation question to calculate their expected GPA.

Secondary Outcomes to study the mechanisms of the treatment effect

Variables used as outcomes to study the channels through which the treatment has an effect on students' preferences and behavior.

- *Interest in studies*. Measured in two different ways. First, the difference between the reported interest when studying for the courses in the first survey (before they know whether they have an incentive goal) vs in the second survey (when they have had an

[1] We have noticed from historical data that these courses are the most homogeneous ones. There are very mild group-specific shocks, so this measure can be useful as a secondary outcome when performing the between group analyses (see the *Main Analysis* section).

[2] We are unsure whether we will be able to also get these data.

incentivized goal). Second, the self-reported change of interest from the first to the second semester.

- *Self-reported happiness for each grade*. Difference between the reported happiness in the first vs in the second survey.
- *Self-control*: Self-reported measure in the second survey on whether they have managed to follow their study plans.
- *Willingness to pay to have the goal in the following semester:* Difference between the answer in the first vs the second survey.
- *Percentage of attended lectures*: Difference between the answer in the first vs the second survey.
- *Total number of study hours during one week of the semester:* Difference between the answer in the first vs the second survey.
- *Total number of study hours during the week before the exams*: Difference between the answer in the first vs the second survey.
- *Productivity while studying*: Self-reported measure in the second survey on whether they have felt more focused while studying.
- *Frequency of studying with friends*: Difference between the answer in the first vs the second survey.
- *Number of courses in which they took the continuous evaluation* (compared to their belief in February).
- *Number of courses in which they took the re-exam* (compared to their belief in February).
- *Variance within each subject of the grades in the different courses:* To study changes in studying strategies.

The first four variables allow us to study changes in preferences due to treatment, while the last eight allow us to study changes in behavior.

Experimental Design

The student body that we recruit our participants from consists of first-year students at the Business Administration program of the University of Barcelona. We will approach student by start of the spring term (i.e. their second semester at the University). Students are divided into 11 classes with roughly 80 students per class. In connection to one of their lectures, we will invite each class of students to follow us to a computer room. Each student will be placed in front of a computer and we will provide a link to a web-based survey (administered by Qualtrics). Students will then read a text that invites them to participate in our study. If they participate, they will have to fill out a questionnaire during the following 25 minutes and fill out a survey in July. Everybody who does so will be paid 25€. Furthermore, some of those who fill out the initial questionnaire will be offered 300€ for reaching a goal GPA.

The questionnaire asks questions about their background (gender, age, spending, ...), study patterns (hours they study per week, lectures attendance rate, ...), preferences (self-control, discounting, risk-aversion, willingness to pay for the incentivized GPA goal, …), and beliefs (about their own grades with or without incentives and about others' grades). The questions regarding their preferences and beliefs are incentivized (using Multiple Pricing Lists and belief elicitation methods). We let them know that 20 randomly selected students from one of the eleven classes will be paid according to their choices.

We will repeat this recruitment procedure once/twice a day during one week, until all students of the 11 classes have been invited to participate.

Finally, we will send an e-mail to all the students of the cohort inviting them to participate by filling out the online survey on their own. This enables students who did not attend the lectures (or were unable to visit our computer room sessions) to participate in the study.

An English translation of the questionnaire is available in a separate document at the AEA RCT Registry site.

Once we have a list of all the study participants, the Administrative Office at the university will send us the entrance exam grade and the first semester grades of each participant. Based on this information and the questionnaire's information, we will use an algorithm to randomize students into being treated or not. See "Randomization Method" for more details on how the algorithm works.

Three weeks after the beginning of the experiment, we will send an e-mail to all the study participants informing them about whether they have been assigned the goal or not. During the following two weeks we will also send a receipt of their right to be paid 25€ if they fill out the survey in July. In addition, those who are treated will also be told in the receipt that they are entitled to 300€ if their second semester GPA is equal or above 6 points.

In July, once the re-take exams are over, we will send a second survey to all study participants. That survey will mostly ask similar questions to the survey in February and will mainly be used to study treatment effects. Furthermore, it will ask students for the name of their friends. We will use these data to further study treatment spillovers.

An English translation of the follow questionnaire is available in a separate document at the AEA RCT Registry site.

Randomization Method

Randomization will be done in office by a computer algorithm.

We will randomize both clusters (classes) and individuals (students). We will treat 30% of the students in 6 classes and 70% of the students in 5 classes.

The treatment allocation algorithm starts by a computer selecting one class that is the most extreme in terms of first semester GPA. More specifically, for each class we calculate the difference between the average GPA in that class and the average GPA across the remaining 10 classes. We then choose the class with the largest difference.[3] In this class, 30% of the students will be treated. The assignment to treatment will be stratified in 6 strata based on the students' first semester GPA. In addition, 20 students from this class, also selected randomly, will receive one of their incentivized decisions from the questionnaire.

From the remaining 10 classes, 5 will be assigned to a *70-treatment group* and 5 to a *30-treatment group*, with the restriction that each group should contain the same (or as similar as possible) number of classes with lectures in the morning (up to 7 classes) and classes with lectures in the afternoon (up to 4 classes). Out of the possible combinations, we will select those that satisfy the balancing condition that the first semester GPA difference in both groups is less

[3] According to the provisional grades that we have seen, we expect this class to be either B1 or B6, both morning groups, which seem to be the most outlier classes.

than 0.2 points (about 10% standard deviation) and the difference between the number of students in both groups is less than 40 students.

In the 70-treatment classes, 70% of the students will be randomly assigned to the treatment group. The assignment to treatment will be stratified in 6 strata based on the students' first semester GPA.[4] In the 30-treatment classes, 30% of the students will be randomly assigned to the treatment group on the same basis.

To check for balance, we will test that the covariates among the treatment and control groups are similar. We chose these covariates based on how important we think that the interaction between the treatment effect and the covariates might be. In particular, we will check that the following holds:

1. The absolute difference between the number of participants in the control and in the treatment group is lower than 40.
2. The absolute difference between the number of participants *who filled out the questionnaire in the lab* in the control and in the treatment group is lower than 20.
3. The p-values of the following t-tests comparing students in the control group with those in the treatment group are higher than 0.6: first semester GPA and entrance exam grade.
4. The p-values of the following t-tests comparing students in the control group with those in the treatment group are higher than 0.3:
    - first semester GPA for 2 different ability groups (according to their first semester GPA),
    - first semester GPA for 3 different ability groups (according to their first semester GPA),
    - gender,
    - spending,
    - age,
    - how often they study with other colleagues,
    - self-control score (following the Tangeny et al. 2004 questionnaire),
    - interest in studies,
    - whether they think that they will study more with the goal,
    - whether they believe that they would be able to study more than currently,
    - WTP for the goal,
    - expected grade improvement with the goal,
    - estimated time discounting, δ, from a time preference elicitation task
    - estimated present bias, β, from a time preference elicitation task

If one of these four conditions is not satisfied, then the algorithm will redo the treatment assignment and check these conditions again. We will run such rerandomization until one assignment satisfies the four conditions.[5]

[4] The stratification is made separately for each class.

[5] Whether these conditions are too restrictive or too permissive will depend on the number of subjects and the correlations between the different variables. We aim for about 0.5-2% of the randomizations to pass this check. We do not aim for a lower percentage because of computational restrictions when running the analyses (for example, if only 0.1% of the randomizations passed the check, we would need about 2000 hours of simulations to run the rerandomized analyses (see the *Power analysis* section)). We do not aim for a higher percentage because we would then lose the potential to have a more balanced experiment. Thus, if the number of randomizations that

Planned Number of Clusters

11 class groups.

Planned Number of Observations

About 600 students.

This number does however depend on the number of students that decide to participate, that can be any number between 400-800. The university administrative staff have told us that they think that about 600 students will participate. One of the conditions that we had to agree with to run this experiment was that all students should be offered the possibility to participate.

Sample size (or number of clusters) by treatment arms

About 280 students in the treatment and 320 students in the control group (if the sample size is 600).

Power calculation: Minimum Detectable Effect Size for Main Outcomes

We used data from the academic year prior to our study (2016-2017) to perform a power analysis with a sample of 600 students. The power analysis is not done on the basis of a t-test, but rather a regression analysis in which we explain students' second semester GPA with their grades in the first semester, their class group, and their treatment condition. Because in the analysis we will also control for their entrance exam grade (that we did not have for the academic year 2016-2017) and multiple covariates (age, gender, spending, study habits…) we expect that we have power to detect treatment effects that are smaller than ones we report here.

We estimate that we have 80% power to find a treatment effect on the students' second semester GPA of 0.21/10 points. Last year's average second semester GPA had a mean of 4.99 and a standard deviation of 1.84. Thus, we have 80% power to find an effect of about 11% standard deviation.

See more information in the section *Power Analysis*.

Main analyses

We will perform the analyses in two ways. The first one will be using the traditional parametric regressions. While the estimates in these analyses are not biased, the p-values will be too conservative. This is because such analyses do not account for the rerandomization method that we used to assign students to the treatment and to the control group. The second analysis will account for this problem by simulating 10.000 treatment assignments that would have passed our balance check. We will then use these 10.000 assignments to more accurately estimate the p-values.

We will divide the analyses into the following categories.

Treatment effect on students' academic outcomes. We will use parametric regressions to explain students' academic outcomes (the ones described in the above outcome variables section) using their grades in the first semester, their entrance exam grades, their class group, their treatment assignment, and the covariates (age, gender, spending, study habits…) that we obtain through

pass these checks is not between 0.5-2%, we will establish (and register) a new rule that satisfies this condition and then pick the first treatment assignment that satisfies such rule.

the questionnaire. We will test if the treatment effect is different between ability groups. We will run this analysis both based on two ability groups and three ability groups. We will also perform exploratory analyses where we will interact the treatment condition with the different covariates to explore which students are most affected by the treatment effect.

Peer-effects on students' academic outcomes. We will use parametric regressions to compare the grades of the students who are in the classes where 30% were treated versus the grades of the students in the classes where 70% were treated. First, we will analyze the peer effect for the entire sample, assuming that for any student (whether treated or untreated) there is a constant effect on having a higher proportion of class colleagues who are treated. Then, we will perform the same analysis but differentiate between students in the control group and the treatment group. The analyses will be the same as in the previous point, but this time without class fixed effects. In these analyses, we will also control for the percentage of total participants (with respect to the number of class students) in each class interacted with whether the group has 70% or 30% treated participants. This will control for the actual percentage of treated students in class (note that probably not all students will be participants).

Treatment effect on students' preferences and habits. As described in the *Secondary Outcomes* section, we will study whether the treatment has any impact on students' self-reported study behaviors and preferences.[6]

Students' characteristics on their goal valuation. We will explore whether there are characteristics of the students (such as their self-control, interest in studies, reference point, and study habits) that predict their WTP and their non-monetary goal valuation (see *Secondary Outcome*).

Students' goal valuation on the treatment effect. We will study whether students who value more the goal and who believe that the goal will have a larger impact in their grades are also those who display larger treatment effects. This will speak on whether students are aware of the effects that such goals will have on them.

Welfare analyses. We will perform a welfare analysis of the policy of paying 300€ to reach a GPA of 6 points based on students' beliefs, WTP and actual outcomes. To do so, we will assume that students only like the policy because it helps them to get better grades, and estimate using their beliefs how much students are willing to pay to obtain better grades. We will then use their actual outcomes to estimate their monetary utility of the policy. We will perform further assumptions to extrapolate the analysis to the policy of getting paid 300€ to reach a GPA of 8 points (for which we ask for their WTP). Here we leave open some degrees of freedom to perform further welfare analyses.

Network analysis. We will create a network based on the participants' answers in the last survey, where we ask them for their friends. First, we will analyze whether having friends who are treated affects their performance. Second, we will explore different ways to analyze the effects of incentives on such networks. We will only perform this analysis if the data about the participants' friends obtained in the last survey is sufficiently rich (that is, that most students answer it).

[6] We will further try to collect data on their class attendance, but we are unsure whether we will able to do it.

Hypotheses

We hypothesize that the treatment effect will be positive in most of the outcome variables that we have described. We believe that students under the treatment will obtain higher second semester GPA, get higher pass rates, improve their second year grades and improve their study habits. Furthermore, we believe that treated students' reference point will increase, self-control problems will decrease and interest in their studies will also increase.

Although we believe that the average treatment effect will be possible, we consider the possibility that in different parts of the distribution (for example for the top achievers) the treatment effects can be null or even negative.

We hypothesize that the treatment effect will be higher for those students who have low intrinsic motivation and for those who report having self-control problems.

We hypothesize that on average the students' non-monetary goal value will be positive. Furthermore, students with higher non-monetary goal value will be more positively affected by the treatment effect.

For the peer-effects analysis, we do not have any clear hypotheses.

Power analysis

We used data from the academic year 2016-2017 to perform a power analysis. These data consist of the grades of every first year Business student in each course during the first two semesters of studies. Because the outcome of our experiment is students' second semester GPA, we have had the possibility to study how their first semester grades explain their second semester GPA and the treatment effects that we have power to detect.

For the power analysis, we randomly selected 607 students (about 80% of the students in the morning groups and 60% in the afternoon groups, based on several professors' predictions of the likelihood that we would recruit students for the experiment) and assigned each of them to a placebo treatment group or to a control group. We did that such that in 6 classes 30% of the students were treated and in 5 classes 70% of the students were treated. We assigned students to each condition in three main different ways: complete randomization, algorithm randomization, and algorithm rerandomization. Complete randomization assigned students randomly. Algorithm randomization assigned students according to an algorithm that mimics the algorithm proposed above to the extent that it is possible given that that all the variables are not available in old data. In particular, the randomization algorithm made sure that:

1. The difference between the number of participants in the 30% classes vs. in the 70% classes is lower than 50.
2. The difference between the first semester GPA in the 30% classes vs. in the 70% classes is lower than 0.2/10 points.
3. Within each of the 11 groups, students are assigned to either the treatment or control according to 6 strata based on their first semester GPA.

Algorithm rerandomization uses the algorithm randomization assignments and only picks those that satisfy several conditions. In this analysis, the conditions are that:

1. The p-value of the students' first semester GPA t-test between treated and untreated students is over 0.6.

2. The p-value of the treated students' first semester GPA t-test between the 30% and 70% classes is over 0.6.
3. The p-value of the untreated students' first semester GPA t-test between the 30% and 70% classes is over 0.6.
4. The p-value of the students' first semester GPA t-test between treated and untreated students is over 0.3 for the following ability categories: bottom half, top half, bottom third, mid third and top third.

About 8% of the algorithm randomization assignments satisfy these conditions. Notice that in the experiment we will have additional conditions, such as gender, age, study habits, etc. (see section Randomization Method above for more details).

Once students are assigned to groups (with whatever of the three randomization methods), we perform the same statistical analyses to explain their second semester GPA, as we will do in the experiment. For each randomization method, we repeat the process 10.000 times. By storing the placebo treatment effect of each simulation, we obtain a distribution of possible effect sizes. Notice that if we assume (as conventionally) that the treatment effect is constant for all treated subjects, it becomes straightforward to calculate the effect size that we have power to detect.

We report the results of the simulations for different randomization methods and tests in Table 1 and 2. The reported treatment effects (TE) are measured in the grading scale of 10 points and describe the effect sizes needed to have 80% power to reject the null hypothesis of no treatment effect using a 5% significance level. The treatment effects in Table 1 (within analyses), show that using algorithm rerandomization we can detect treatment effects that are 10-50% smaller than using complete randomization. In the peer effects tests in Table 2 (between analyses), we show that the gain is greater than 50%. Figures1 and 2 represent such gains in power by showing that the distribution of placebo treatment effects is less spread when using the algorithm rerandomization method to assign students to treatment.

Table 1. Effect size required to detect 5% significance with 80% power by randomization method (lower is better). Within class analysis.

	Sides statistical test	TE whole sample	Heterogeneous treatment effects				
			Two ability levels		Three ability levels		
			TE (bottom half)	Interaction T*top half	TE (bottom third)	Interaction T*mid third	Interaction T*top third
Complete randomization	Two	.2403	.3950259	.4586105	.5051435	.6639005	.5684023
	One	.2154	.3516363	.4065815	.4475238	.5882617	.5012354
Algorithm randomization	Two	.2339	.3583724	.425613	.3992048	.3910412	.4698404
	One	.2078	.3196647	.3793517	.3554348	.350244	.414489
Algorithm rerandomization	Two	.2104	.2835413	.308188	.3331355	.3094301	.3476053
	One	.1881	.2666086	.2978924	.3089505	.2974165	.3325668

Note. This table reports the treatment effect (TE) (in the grading scale of 10 points) for which the experiment has 80% power to detect with 5% significance, depending on the test and on the randomization method used. Such effect size has been calculated through 10.000 simulations of OLS class fixed-effects regressions explaining students' second semester GPA with their first semester grades. The sample are 607 students randomly selected from the academic year 2016-2017.

The three randomization methods create 6 classes with 30% treated students and 5 classes with 70% treated students. *Complete randomization* means randomizing without any condition. *Algorithm randomization* means randomizing with the conditions that the six 30% classes are similar to the five 70% ones (in terms of GPA and number of participants) and 6 strata according to students' GPA within each group. *Algorithm rerandomization* picks only allocations to treatment from Algorithm randomization that satisfy several balancing conditions (based on students' GPA, students' GPA for different ability levels, and students' GPA in the control/treatment group in the 30% classes and in the 70% classes). In these concrete tests, about 8% of the algorithm randomizations pass the balancing conditions.

Sides statistical test reports whether the test is one-sided or two-sided. *TE whole sample* is the average treatment effect required to find significance with 80% power across all students assigned to control and treatment group. The *Heterogeneous treatment effects* shows the required treatment effect for the baseline ability level (bottom half or bottom third) and the size of the additional treatment effect needed for the other ability levels to be significantly different from the baseline category.

Table 2. Effect size required to detect 5% significance with 80% power by randomization method (lower is better). Between class analysis.

	Sides statistical test	Baseline Effect (of being in the 70% groups).	Additional effect on treated T*I{70%}
Complete randomization	Two	.7517073	.5622784
	One	.6806058	.4946123
Algorithm randomization	Two	.7062734	.486405
	One	.6509027	.4373381
Algorithm rerandomization	Two	.3224563	.2443547
	One	.2925724	.1646692

Note. The details of this table are the same as in Table 1. Here, however, the regressions do not contain class fixed-effects. Instead, we drop from the analysis the outlier 30% class (see Randomization Method) and include a dummy that indicates if the class has a treatment fraction of 70%. The *Overall effect* is the required coefficient of such dummy for the entire sample (this is, it assumes a constant effect, for both treated and untreated students, of being in a 70% class instead of a 30% class). *Additional* e*ffect on treated* shows the effect size required for the interaction between treatment (i.e. incentivized study goals) and being in a 70% class.

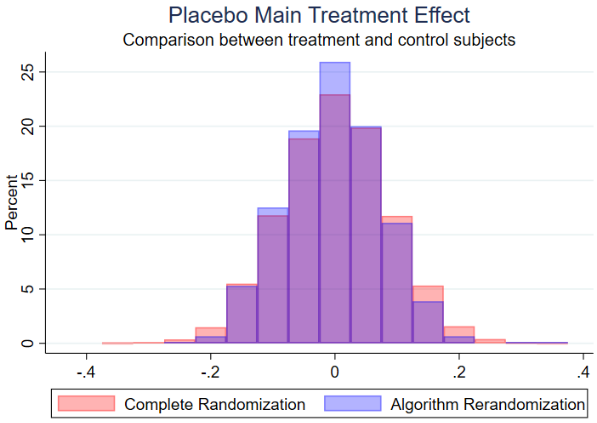

Figure 1. Distribution of the placebo treatment effect (comparison of a control and a treatment group when there is no such effect) using an OLS class fixed-effects with multiple controls that explains students' second semester grades. The x-axis measures the treatment effect in the grading scale of 0-10 points The analysis is based on 10.000 simulations with the grades of 607 students of the academic year 2016-2017.

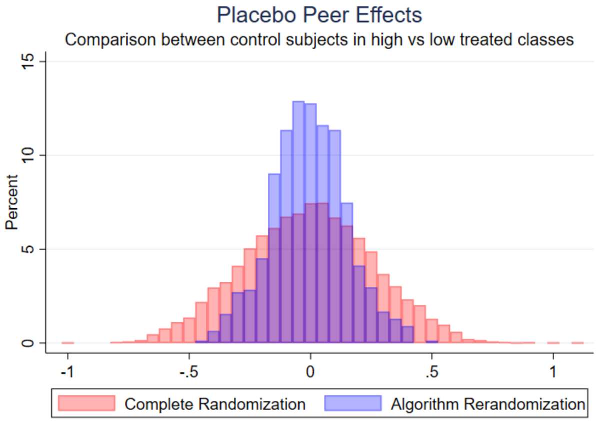

Figure 2. Similar to Figure 1 this is the distribution of the placebo treatment effect when comparing untreated students in the 70% classes with untreated students in the 30% classes.

There are a few noteworthy issues with the power analysis. First, while power analyses typically study the sample size needed to detect a *reasonable* effect, we fix the sample size and study how large the treatment effect needs to be in order for us to detect it. This is because our agreement with the university states that we have to invite all students to participate in this experiment, so we will not be able to choose our sample size. We have thus decided to fix the sample size and study whether the treatment effects that we have power to detect are reasonable. Such analysis has been crucial both for deciding how to design the experiment and for choosing what analyses to pre-register.

Second, the analysis here presented is overly conservative. Notice that in the actual experiment we will be able to control for many additional variables, such as students' grade in the university entrance exam, gender, age, spending, class attendance, study hours, etc. These control variables will increase the explanatory power of the regressions, implying that we will have power to detect even smaller treatment effects. According to our estimations (not reported), by controlling for such variables we will be able to further reduce the required effect size to have 80% power by about 10-30%.

Third, we also checked the OLS regression p-values of the randomizations that generated the effect sizes in the 97.5th percentile (i.e. those randomizations that define the lower bounds in Table 1 and 2). While it is around 0.05 in the complete randomization in Table 1, it becomes consistently above 0.05 for algorithm rerandomization. This shows (as expected) that using traditional regression methods to analyze an assignment made through rerandomization will yield overly conservative p-values.

Fourth, we note that we have relatively weak power to detect peer-effects (Table 2). Previous research on incentives for university students have found effects of about 20% standard deviation (see for example Leuven et al. 2010 and de Paola et al. 2012). In our experiment, this would be a treatment effect of about 0.36 points (in a scale from 0 to 10). If this was our treatment effect, note from Table 2 that to have 80% power, the required peer effect would have to constitute 67% of the overall incentive effect. Peer effects that size may sound unrealistic. However, we believe that we will considerably improve this power. First, we will be able to control for a rich set of covariates that will help us to narrow down such peer-effects in a much better way. Second, we will use data from the teachers in the previous year to get a measure of how much each teacher contributes to the second semester GPA of each student (for example, by how "good" the teacher is). By using these data for the year during which we run the experiment, we hope to control for a substantial part of the shocks at the class level that we see in our sample. Third, previous literature has suggested that such peer-effects can be either positive or negative. Even if in the end we do not reach a reasonable power level to capture this effect (that we hope we do), we believe that we will be able to contribute to the literature regarding the size and sign of such effect. Finally, we considered treating 50% of the students in all classes. This design did however leave us completely blind with respect to any potential peer-effects and did almost not contribute to improving our power. Using the *algorithm randomization* method, we simulated and compared the effect size that we could detect with 80% power when incentivizing 50% of the students in each class vs. when incentivizing 30% in 6 classes and 70% in 5 classes. The effect size required to have 80% power is only 4-8% lower in the 50% design across all tests. We believe that this gain is very small compared to the loss of not learning anything about the peer-effects results of this policy.

Finally, note that we have done the analysis with 607 randomly selected students. However, we do not know how many students we will be able to recruit. To be sure that our power does not fall down to unreasonable levels if less students decide to participate, we performed the same analyses with a total of 437 randomly selected students. The power does obviously fall, but not to unreasonable levels. For example, using the algorithm rerandomization method, we have 80% power to detect a treatment effect for the whole sample of 0.2557 (instead of .2104). The proportions are similar across all the different tests: we need effect sizes about 20% larger with 437 participants than with 607 students to have 80% power to detect them.

References

Leuven, E., Oosterbeek, H., & Klaauw, B. (2010). The effect of finacnial rewards on students' achievement: evidence from a randomized experiment. Journal of the European Economic Association, 8(6), 1243-1265.

De Paola, M., Scoppa, V., & Nisticò, R. (2012). Monetary incentives and student achievement in a depressed labor market: results from a randomized experiment. Journal of Human Capital, 6(1), 56-85.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. Journal of personality, 72(2), 271-324.