Essays in Education and Health Economics

Cristina Adelaida Bellés Obrero

TESI DOCTORAL UPF / ANY 2017

DIRECTORS DE LA TESI
Dr. Robin Hogarth i Dr. Marc Vorsatz

Departament d'Economia i Empresa

upf. Universitat Pompeu Fabra
Barcelona

A mi padres y mi hermana.

Acknowledgements

The completion of this dissertation would not have been possible without the invaluable academic and emotional support of numerous people, to only some of whom it is possible to thank here.

First and foremost, I would like to express my sincere gratitude to my advisors Robin Hogarth and Marc Vorsatz for their guidance and thoughtful encouragement through all these years. This Thesis would have never taken shape without their continuous support, and comments.

I am particularly indebted to Judit Vall and Sergi Jimenez for their excellent guidance, care, and patience. Working with them was not only a very productive and stimulating experience, but also a source of motivation. They were exceptionally generous with their time and I am extremely grateful for having their office doors always open to me.

In addition, I gratefully acknowledge all the insightful comments and suggestions from all the members of the applied microeconomics and the behavioral groups at UPF, with special mention to Libertad Gonzalez and Gianmarco León. I also owe many thanks to all the members of the Center for Research in Health and Economics at UPF. My special thanks to Jose Apesteguia whose initial conversations and advice were key in shaping my way of thinking about research. I would also like to extend my deepest thanks to Pedro Rey-Biel and Antonio Cabrales for their invaluable comments on my papers.

I would also like to thank Marta Araque and Laura Agusti for their assistance, especially during the job market, and for making all this process easier for us.

I had the great opportunity and good fortune of sharing this path with other students that have become great friends. I would like to specially thank Ana, Darren, Francisco, Helena, Lorenzo, Marco, María, Marta, Miguel, Niklas, Ricardo, Thomas, and Tomás for the countless coffee breaks, beers, dinners, barbecues,

and trips. I am sure my PhD experience would have not been the same without all of you. I also owe a lot to many other friends that have bore me while doing the PhD. I am especially grateful to “mis Ouets” (Almu, Joan and Neus) , Chiara, Shree, and all my friends from Benafigos.

Finally, I would like to express my profound gratitude to my parents, Jose Manuel and Juana, and my sister Marta. They had an unshakable faith in my success and were by my side in all the good and bad times. This Thesis would not have been possible without them. Thank you.

Abstract

This dissertation consists of three chapters that investigate students' and teachers' incentive programs, and the intergenerational infant health consequences of a labor market policy. In the first chapter, I perform a randomized control trial at a distance learning university to compare three different monetary incentive schemes varying students' performance target in the same educational environment. I show that the performance target implemented interacts with some of the characteristics of the students incentivized, such as intrinsic motivation and experience with the incentivized task. Moreover, a novel finding of this study is that incentives foster students' strategic behavior that is triggered by the way performance is measured. In the second chapter, I examine how tying teachers' pay to students' performance affects the latter's achievements. I show that a nationwide program implemented in Peru giving monetary rewards to teachers conditional on their students' performance, has a precisely estimated zero impact on students' grades. Finally, in the third chapter I investigate the effect of a child labor regulation that increased the minimum legal age to work from 14 to 16 years old, on fertility and infant health outcomes. Using a difference-in-differences strategy, I find that the reform increased educational attainment, and decreased marriage and fertility. Interestingly, I show that the reform was detrimental for the health of the offspring at the moment of delivery.

Resum

Aquesta tesi s'estructura en tres capítols que investiguen els programes d'incentius per als estudiants i professors, i les conseqüències intergeneracionals per a la salut infantil d' una política de mercat de treball. En el primer capítol vaig duu a terme un experiment de camp en una universitat d'educació a distància amb la finalitatat de comparar tres incentius monetaris diferents en el mateix entorn educatiu, variant l'objectiu de rendiment dels estudiants incentivat. Mostro que l'objectiu de rendiment implementat interactua amb algunes de les característiques dels estudiants, com ara la seva motivació intrínseca i l'experiència que tenen amb la tasca incentivada. D'altra banda, també trobo que els incentius fomenten el comportament estratègic dels estudiants com a conseqüència de la manera en la que es mesura el seu rendiment. En el segon capítol examino com afecta a l'assoliment dels estudiants el fet de que la retribució dels seus professors estigui lligada al seu rendiment acadèmic. A aquests efectes, analitzo un programa nacional implementat a Perú que dóna una recompensa monetària als mestres condicionada al rendiment dels seus alumnes. El programa té un efecte nul precisament estimat sobre les qualificacions dels estudiants. Finalment, en el tercer capítol investigo l'efecte sobre la fertilitat i la salut del canvi legislatiu que va augmentar l'edat mínima legal per treballar de 14 a 16 anys. Utilitzant una estratègia de diferències en diferències, arribo a la conclusió que la reforma va incrementar el nivell d'educació, alhora que va disminuir la fertilitat i probabilitat de contraure matrimoni. Addicionalment, mostro que la reforma va ser perjudicial per a la salut de la descendència en el moment del part.

Preface

This doctoral thesis brings together three independent research projects that involve the study of incentive programs in the educational context, and the empirical analysis of a labor market policy over fertility and infant health outcomes.

Even though there are substantial socioeconomic returns to education, underinvestment and low educational performance are still important problems in many developed and developing countries. As a consequence, many governments and policy makers have implemented monetary incentive programs, targeting teachers and students, to boost educational attainment. Students' incentive programs aim to correct for students' insufficient motivation, excessive discount rates for the future, or underestimation of the returns to education, which are considered the be main reasons for their suboptimal study effort. Given that teacher quality is a key factor determining student achievement, teachers' incentive programs, on the other hand, intent to compensate for the incentives generated by teachers' compensation policies (flat salary progression and lifetime job tenure) that create weak incentives for teachers to exert high levels of efforts. However, evidence on the effectiveness of these programs is still scant and inconclusive. Moreover, previous studies have primarily focused on the question of whether a particular type of incentive works in a specific educational environment, overlooking the role of the incentive design. The first two chapters of this thesis aim at shedding some light on the impact of these incentive programs and their design on students' attainment.

In the first chapter, I investigate which features of these incentive programs are effective at increasing students' attainment and how these interact with the characteristics of the students that are being incentivized. For this purpose, I conduct a randomized control trial at a distance learning university in Spain to compare three monetary incentive schemes with different performance targets for students. The distance learning setting is particularly advantageous for this intervention as it guarantees independent observations (no spillovers between treatments and the control group) and no effect on teachers, at a very low cost. The first treatment (Threshold) provides a cash reward for students who achieve a grade threshold, the second (Top percentile) for students in the top of their class, and the third (Improvement) for those that improve their expected grade. From a theoretical point of view, incentive schemes with different performance targets should increase attainment of students with different characteristics. I focus on the interaction of the incentive schemes with three characteristics of students that have been widely discussed in the literature on education as determinant for incentives to work: intrinsic motivation, experience with the incentivized task, and ability. I find that for the "Top percentile" treatment, the effect of the monetary reward is positive for students with a high intrinsic motivation and negative for students with a low intrinsic motivation. On the other hand, the "Threshold" and "Improvement" incentives have positive effects on students with more experience with the incentivized task and negative effects on those with less experience. Finally, a novel finding of this paper is that incentives foster students' strategic behavior, that is triggered by the way performance is measured (multiple choice exam with penalties for incorrect answers). The study emphasizes the importance of incentive design. The question of whether monetary incentives are effective in increasing students' attainment is too narrow and further research is needed to identify the features of incentive design that matter in practice as well as how different design features interact.

In the second chapter, María Lombardi and I focus on incentive programs for teachers. Specifically, we investigate how tying teachers' pay to the performance of students affects the latters' achievements. We conduct an evaluation of "Bono Escuela", a nationwide program implemented in public secondary schools in Peru that gives monetary rewards to teachers conditional on their students' performance in a standardized test. The program takes the form of a tournament, awarding a bonus of over a month's salary to the principal and every teacher from schools in the top 20 percent within a group of comparable schools. The main measure used to rank schools in the tournament is the average score of 8th graders at the standardized test. Using a novel administrative database covering the universe of Peruvian students, we perform a difference-in-differences estimation comparing the change in the internal grades of 8th graders before and after the incentive was introduced to that of 9th graders from the same school. We find that the program had a precisely estimated zero impact on students' grades. We argue that this zero effect can be explained by some aspects of the program's design (students' low stakes, teachers' inexperience with the standardized test, and the program's timing), which may have hindered teachers' ability to improve the incentivized outcome or infer their probability of winning. The study sheds some light over the scarce literature on scaled-up teachers pay-for-performance programs in developing countries and highlights the need of a deeper understanding about the role played by the different characteristics of these programs in their success.

My third chapter, co-authored with Sergi Jiménez Martín and Judit Vall Castelló, studies the effect of a child labor regulation on fertility and infant health outcomes. For this purpose, we examine a child labor reform that took place in Spain in 1980, which increased the minimum legal age to work from 14 to 16 years old while the compulsory school age was maintained at 14. Contrary to previous literature, we exploit the interaction between the compulsory schooling age and the minimum legal age to work, as both age thresholds are important to determine an individual's decision to remain in the educational system. We perform a difference-in-differences strategy to identify the reform's within-cohort effects, where treated and control individuals only differ in their month of birth. We find that the reform increased educational attainment, but decreased womens' completed fertility and reduced their probability of ever marrying. In addition, we show that the reform was detrimental for the health of the offspring at the moment of delivery. Newborns of women affected by the reform had a higher probability of being premature and having low birth weight. We identify three different channels contributing to this detrimental effect: the postponement of fertility, the change in the maternal marital status, and the improvement in the labor market conditions of more educated women, which increases the likelihood of engaging in unhealthy behaviors such as smoking for the affected cohorts. This last surprising result is a direct consequence of the socioeconomic situation of Spain when the reform took place. Women in these cohorts grew up during the early post-Franco era and experienced the process of gender equalization. As a consequence, these more educated women had more access to and social acceptance of smoking compared with pre-reform cohorts. This study is informative, from a policy perspective, for developing countries whose educational system, child labor market participation rates, and womens' social development are similar to the levels that Spain was experiencing around 1980.

Contents

List of Figures

List of Tables

Chapter 1

Who is Learning? A Field Experiment Comparing Three Different Incentive Schemes in the Same Educational Setting

1.1 Introduction

There are substantial socioeconomic returns to college education. College graduates have higher future earnings than high school graduates (Kane and Rouse, 1995; Zimmerman, 2014; Angrist and Chen, 2011). They also enjoy better health, live longer (Buckles et al., 2014), and participate more in civic life (Dee, 2004). Low degree completion and elongation of time to degree[1] are still important problems in many developing and developed countries. Students may exert suboptimal study effort if they have excessive discount rates[2] (Oreopoulos, 2007), time-inconsistent preferences[3] (Cadena and Keys, 2015), or they underestimate

[1] For instance, 45% of college students in the US (Shapiro et al., 2012) and 25% in Canada (Shaienks and Gluszynski, 2007) fail to complete a college degree within six years. Spain also suffers from high drop out incidence and long duration of the studies. 19% of college students in Spain drop out during the first year of university (Ministerio de Educación, 2015), and almost 50% of students require two or more additional years to finish their studies (Lassibille and Gómez, 2011).

[2] Previous literature has found that children and adolescents exhibit high discount rates and have difficulties to plan for the future (Bettinger and Slonim, 2007; Steinberg et al., 2009).

[3] People tend to have hyperbolic time preferences, overweighting the present and largely ignoring future rewards, inducing dynamically inconsistent preferences (Laibson, 1997). Previous

the returns from education (Eckstein and Wolpin, 1999). As a consequence, governments, policymakers, and universities have devoted significant resources to investigate if monetary incentives can boost educational attainment and increase post-secondary on-time completion rates. For instance, Georgia's Helping Outstanding Pupils Educationally (HOPE) scholarship program, that started in 1993, pays for tuition, fees, and books at any public university for students with a high school GPA of at least a B as long as they maintain the same grade in college. The popularity of this program has spread and nowadays other states have similar merit scholarship programs. Similarly in Spain, students that graduate from high school with honors do not pay first year's tuition at any public university.

Yet, the effect of economic incentives on students' performance is far from unequivocal (Gneezy et al., 2011). While some studies show positive effects (Kremer et al., 2009; De Paola et al., 2012; Levitt et al., 2016b), many show positive results for some groups and no effects for others (Angrist and Lavy, 2009; Angrist et al., 2009; Bettinger, 2012; Angrist et al., 2014), or no effects at all (Fryer, 2011; Barrow and Rouse, 2013).

Given these mixed results, it is appropriate to ask which features of incentive programs are effective at influencing students' attainment. Previous studies have focused primarily on whether a particular type of incentive works in a specific educational environment and has overlooked the role of incentive design.[4] These studies differ not only in the type of incentive scheme used but also in the economic and social context where the experiments are conducted, the duration, or size of rewards. It is thus problematic to compare effects of different incentive schemes from the available field experiments.

The present paper contributes to the existing literature by comparing three incentive schemes that vary students' performance target in the same educational environment. In doing so, we seek to answer the following question: Which

studies have stated a negative correlation between hyperbolic discount rates and educational outcomes (Kirby et al., 2005).

[4] Some studies have varied the reward recipient (Behrman et al., 2015; Levitt et al., 2016a), framing (Levitt et al., 2016b), task (Hirshleifer, 2015), and payment structure (Levitt et al., 2016a).

types of incentive schemes work in education and for whom? This question is particularly important given that, from a theoretical point of view, incentives with different performance targets should boost attainment of different types of students.

There has been substantial academic interest in incentive based education programs.[5] The existing studies that examine incentives on students' performance consider three different types of economic incentives – "Threshold", "Top percentile", and "Improvement" – at two different levels of education, school and college. Studies that incentivize students who reach a certain grade threshold find no effect on students' grades at all (Fryer, 2011) or an increase in grades for students whose ability is close to the threshold (Angrist and Lavy, 2009; Bettinger, 2012; Leuven et al., 2010; Barrow and Rouse, 2013; Angrist et al., 2014). In studies where the reward is given to students in the top percentile of the grade distribution, there are overall improvements in attainment, but especially for high ability students (Kremer et al., 2009; De Paola et al., 2012). Finally, there are some effects in studies where rewards are given to students that improve their previous grades by a certain amount. For instance, Angrist et al. (2009) finds that this type of incentive especially improves girls' grades that are at the lower and higher quartiles of the grade distribution. On the other hand, Levitt et al. (2016b) find overall positive effects on the standardized test grades.

In this paper, we conduct a randomized control trial that compares these three common incentive schemes at a public distance learning university in Spain, Universidad Nacional de Educación a Distancia (UNED). This particular setting (a distance learning university) has several advantages. First, students in a course typically do not know each other. Thus, in this setting randomization can be done within a class without worries of possible spillovers from treatment to comparison groups that could bias the estimation of treatment effects.[6] Second, this setting

[5] Some related studies analyze the effect of conditional cash transfers in México (Behrman et al., 2005), Colombia (Barrera-Osorio et al., 2008), US (Dee, 2011) or UK (Dearden et al., 2009). Others incentivize post-secondary enrolment in the US (Rodriguez-Planas, 2012; Carrell and Sacerdote, 2013).

[6] For instance, the effect will be upward bias if students in the control group decrease their

allows comparing different economic incentives within one course. This avoids the "teaching-to-the-test" problematic[7], as any change in teachers' behavior will affect the treatment and control groups in the same way. Third, UNED courses are larger than in other universities. In fact, the comparison of the three types of incentives within the same course would be impossible in other settings due to the lack of observations.

The experiment involves 1,079 students enrolled in a class of microeconomics in the second year of the Degree in Business Administration at UNED during the academic years of 2014-15 and 2015-16. At the beginning of each academic year, students are randomly assigned to one of the three treatment groups or the control group. In the first treatment, called "Threshold", all students that achieve a minimum grade of 8 (out of 10) in the final exam of the course enter a lottery that gives 1,000€ to the unique winner. In the second one, "Top percentile", students compete against each other, and the top 25% students enter a lottery for the 1,000€ prize. Finally, the last treatment, "Improvement", students have to increase by 1 point their expected grade (that was previously provided in a questionnaire), in order to enter the lottery. Students are informed by email of their assignment status. In this email, it is explained that the university is going to incentivize the work of the students in that specific course and the type of incentive that they are subject to; however, the existence of the different economic incentives is not divulgated.

We expect that providing students with immediate extrinsic monetary incentives could encourage them to exert more effort, positively impacting their performance. On the contrary, these incentives might yield no improvement if students are too far away from their performance target or do not know how

effort as a consequence of feeling cheated, while the bias can be downward if the students in the control group are afraid of losing ground to the other students and increase their effort.

[7] Some studies, with school-based randomizations where teachers knew the treatment status of their students, observe a change in teachers' behavior in treated schools. Teachers in treated schools might have higher incentives to target topics likely to appear in the incentivized exam, coaching students on test-taking strategies, or even cheating. If this change of teachers' behavior is taking place, it is difficult to disentangle the real impact of the incentive over students' effort and learning.

to increase their achievement. For instance, Fryer (2011) states that one of the reasons why financial incentives had no effect on the incentivized outcomes is that students "had little idea about how to translate their enthusiasm into tangible steps designed to increase their achievement".

Students incentives might also be ineffective, or even detrimental if they lead students to engage in strategic behaviors which the objective of increasing the incentivized outcome but not leading to an increase in learning. For example, Cornwell et al. (2005) find that the Georgia HOPE scholarship reduced by 6 percentage points the fraction of freshmen completing a full course load. This effect was driven by the students' strategic reaction to the program's characteristics. The HOPE scholarship rules encouraged students to reduce course loads in order to raise their GPAs and maintain the scholarship. Finally, student's incentive could also decrease their performance if it crowds out students' intrinsic motivation. Leuven et al. (2010) find that the economic reward had a negative effect over students with low ability and they interpret it as an indication that their intrinsic motivation is reduced by external rewards.

In line with much of the existing literature, we find no average effect for any of the incentives on students' performance. However, we reject two reasons that could prevent us from detecting a significant effect. First, the incentive offered is large enough to compensate for the cost to the students of increasing effort.[8] Second, we have enough statistical power to detect medium to large effects of the economic incentives.[9] And yet, we might not be capturing any positive (or negative) average effect because whereas the economic incentives might improve

[8] The expected value of the incentive (4.5€ per credit) is analogous to the low reward condition of one of the most similar interventions (De Paola et al., 2012) that find positive and significant effects. In addition, we expect the lottery to increase the salience of the incentive due to an overestimation of the low probability of winning by the students. Finally, we could argue that if the amount of the incentive mattered to find average effects, we will expect that those students who are more cash constrained (those that do not work but are willing to work) to be more incentivized. However, we find that the average effects do not differ by the working status of the students.

[9] We have enough power to detect an effect larger or equal to 0.17 standard deviations, which is equivalent to an increase of 0.4 points in the final exam or 7 percentage points in the probability of attending the final exam. Only three previous studies have found a smaller effect ranging 0.1-0.15 standard deviations (De Paola et al., 2012; Bettinger, 2012; Levitt et al., 2016b).

the performance of some types of students, they could have negative effects on others. In fact, there are some changes in the distribution of the exam grades between the control group and the treatment groups that suggest that this could be the case.

One objective of this paper is to investigate how the different performance targets of the incentive programs interact with the characteristics of the students being incentivized. Previous literature has highlighted the roles of intrinsic motivation, experience with the incentivized task, and students' ability. More importantly, we expect these characteristics to impact each incentive scheme differently.[10]

The role of intrinsic motivation has been much discussed. On the one hand, if students lack sufficient intrinsic motivation, incentives for achievement could yield increases in student performance. On the other hand, some argue that financial rewards for students undermine intrinsic motivation and lead to negative outcomes. Which of the two effects will dominate is unknown. We expect highly motivated students under the first and second incentive schemes ("Threshold" and "Top percentile") to perform better than students with the same motivation in the control group. On the other hand, students under the "Improvement" incentive with low motivation and, consequently, low grade expectations, should do better in the final exam compared to similar students in the control group.[11] Indeed, we find that for the "Top percentile" treatment, the effect of the economic incentive is positive for students with high intrinsic motivation and negative for those with low intrinsic motivation.

There are several hypotheses that could explain this differential effect. First, achieving a grade threshold can be perceived to be less challenging and less prestigious than being in the top of your class. Highly motivated students could be more attracted while low motivated students could be more discouraged by this latter target. A second hypothesis is that the "Top percentile" treatment

[10] Our predictions, based on an on-going theoretical framework , are summarized in Figure 1.2.
[11] This prediction is based on the assumption of increasing marginal cost of effort.

offers, in addition to a monetary reward, information about relative performance. This information could be valued positively by highly motivated students and negatively by students with low motivation. This latter hypothesis is consistent with previous literature that finds that providing relative performance feedback enhances students' performance (Tran and Zeckhauser, 2012; Azmat and Iriberri, 2016).

Previous studies argue that students might only have a vague idea of how to increase test scores and consequently have little incentive to increase effort (Fryer, 2011). For instance, Angrist et al. (2014) find positive effects among second-year students and claim that this is because these students have "a better sense of how to improve their grades". In each incentive scheme, students face different uncertainties when choosing their optimal effort. While students in the "Top percentile" treatment need to know the distribution of ability in their group, students in the other two treatments only need to know how much effort is required to achieve a certain grade threshold. These uncertainties might reduce the effect of the incentives if students have difficulties understanding how to reach the performance target. Experience with the task that is incentivized (which we proxy by the number of courses students have passed before) could reduce these uncertainties. Yet, given the differences in information required by the three treatments, we expect this reduction to have a greater effect on the "Threshold" and the "Improvement" incentives. As expected, we find that students with more experience have a higher probability of taking the final exam in the "Threshold" and "Improvement" treatments than students with low experience. This result confirms that the information students have when incentivized is important and that experience can help reduce the uncertainties about the grade production function. As students do not interact in this setting, information on their peers is minimal and experience has no impact on the "Top percentile" incentive condition.

Finally, a novel finding of this paper is that incentives foster strategic behavior of students that is triggered by the way performance is measured. We incentivize the final grade on a multiple-choice exam with penalties for incorrect answers. In this type of task, the decision to answer or not an additional question is a strategic issue. We argue that the marginal utility of answering an additional question differs according to students' ability. High ability students have ex-ante a high probability of entering a lottery for the 1,000€ prize. For these students the disutility of answering an additional question incorrectly is very high, as they risk losing the lottery ticket. As a consequence, these students answer fewer questions with the economic incentive than without it. On the other hand, low ability students are far away from entering the lottery. For these students, the disutility of answering incorrectly is rather low while the expected utility of answering correctly, by chance, is high. Thus, these students answer more questions under the economic incentive. We also show that these effects are driven by those questions that are hard thereby supporting the explanation of strategic behavior. This result supports the contention that strategic behavior induced by the incentives is another reason why incentives do not impact students' exam grades.

This paper makes several contributions. First, the study emphasizes the importance of incentive design. We show that the characteristics of the incentive schemes interact with those of the students being incentivized. However, this paper only constitutes a first step towards creating interventions aimed at enhancing educational attainment that are tailored to students' characteristics and needs. Second, the distance learning setting is particularly advantageous for controlled experimentation with several treatment arms within the same course, enough observations, and without possible spillovers that could bias the results. The proliferation of online learning platforms with numerous students constitutes a great opportunity to implement this type of interventions in the future at a lower cost. Third, we learned that it is important to attend more carefully to the type of task that is being incentivized. Very little research has examined this last aspect and this paper demonstrates that the characteristics of the incentivized tasks need to be taken into account. If the task allows, as in the current situation, for strategic behavior of students, the incentive might foster this type behavior.

In summary, our main conclusion is that the focus on the question of whether incentives work in education is too narrow. We have shown that the effects of incentives depend on how they are designed and for whom. Future research should therefore be designed to test alternative features of incentives in the same setting so that knowledge of what works for whom is improved.

There are also other issues for future research. First, we show that there might be complementarities between the type of monetary incentives and the type of information that students have available or can obtain. For instance, the "Top percentile" treatment might provide students with information about their relative performance that is valued differently by students of diverse ex-ante motivation. Also, experience with the incentivized task is important for the "Threshold" and "Improvement" incentives. Second, it is key to understand further the type of tasks that are more effectively incentivized. Previous literature has already compared the efficiency of incentives for outputs or inputs (Fryer, 2011; Hirshleifer, 2015) and concrete or abstract tasks (Bettinger, 2012). Additional research should compare tasks that can induce strategic behavior, such as multiple-choice exams, with those that do not, such as essay exams.

The article is structured as follows. Section 1.2 describes the experimental design and its implementation. Section 1.3 presents the general and heterogeneous results of the experiment. Finally, Section 1.4 concludes and provides some discussion about potential future research questions.

1.2 Experimental Design and Implementation

The experiment involves 1,079 students enrolled in a class of microeconomics in the second year of the Degree in Business Administration at UNED during the academic years of 2014-15 and 2015-16.

UNED is a public distance learning university whose campus is spread across Faculties, Colleges, and Centers in Madrid. Additionally, it has an extensive network of 61 Associated Centers in the rest of Spain and other countries of Europe, America, and Africa. Each student is attached to an Associated Center that they can choose. The Associated Center provides on-site tutoring, organizes the final examinations, and provides students with support facilities. The distance learning process is run through some printed, audio-visual, and multimedia materials. Students also have online as well as on-site (in the Associate Centers) communication with professors. The only compulsory requirement to pass the courses is an in-person exam at one of the Associate Centers. The final exam is a multiple-choice test with 20 questions that have four possible and exactly one correct answers. First-term courses in UNED start at the beginning of October and have two one-week exam periods, one in the last week of January and another in the second week of February.

Placing the field experiment in a distance learning university has several advantages. First, students in each course do not tend to know each other.[12] Thus, we can perform a within classroom randomization without having the usual concern about potential spillovers that could affect the control group and bias the results. The majority of the literature agrees on the convenience of using school-based randomized trials in this type of evaluations, for exactly this reason (Duflo et al., 2007; Angrist and Lavy, 2009). For instance, those individuals in the control group could feel cheated and decrease their effort if they get to know that their classmates are being incentivized. If this happens, our results will be upward biased. On the other hand, the results will be downward biased if students in the control group increase their effort if they are afraid of losing ground to incentivized students. Yet, the randomized trials used to evaluate financial incentives at university level are normally performed within the same university, as there is no standardized performance test at this level. In those studies, to check for these possible externalities they compare participating students that were assigned to the control group with students attending the same course the previous year (De Paola et al., 2012). Though this technique does not rule out the existence of differences among cohorts that might be canceled by the reactions of students assigned to the control group.

[12] Students that participate in the experiment are very spread across the different centers of UNED. The main source of communication will be the forum that the course has. We controlled that students do not talk about the economic incentive in the forum.

Secondly, change in the teachers' behavior among the treatment and control groups is a concern in school-based randomized trials. Teachers of incentivized students could have higher incentives to "teach-to-the-test". Then, the effect of the incentive over learning could not be disentangled from the effect of the "teaching-to-the-test". In our setting, all students share the same teacher independently of their assignment to the treatment and control groups. So, if there is a change in the teacher's behavior as a consequence of the incentive, it will affect all subjects in the same way. Finally, UNED courses are far bigger than in other universities. The comparison of three types of economic incentives over students within the same course would be impossible in other settings due to the lack of observations.

The experiment is performed during the academic years of 2014-15 and 2015-16. At the beginning of each academic year, students have the opportunity of filling out a questionnaire, which is incentivized. Those who fill it out, see their final grade of the course increased by 0.2 points. This questionnaire serves as an implicit acceptance document for participating in the experiment and enables us to extract some information about the students that is not available in their academic records such as their age, gender, age at which they started their studies in UNED, previous studies, the education of their parents, their working status, and the way they finance their studies. The questionnaire also asks for their goal grade for the course as well as the grade they expect to obtain. This information will be used in order to analyze heterogeneous effects of the different incentive schemes. Figure 1.A1 and Figure 1.A2 show the original questionnaire sent to students, as well as the translated version. It is important to note that the motivation given to the students for the filling of the questionnaire is a fixed increment in the final grade. We collected 1,175 questionnaires from students, that is, 48% of the enrolled students answered the questionnaire. We expect there to be a number of differences between students who choose to take the questionnaire (and, as a consequence, participate in the experiment) and those who choose not to. In fact, in Table 1.A5 of the Appendix we show that students that did not participate in the experiment got, on average, 0.28 points less in the previous courses done before in UNED. This means that students participating in the experiment have, on average more ability than students that did not participate in the students. We also observe that non-participants have 12 percentage points higher probability of taking the final exam compared with students that were part of the experiment and assigned to the control group. On the other hand, those students in the control group that did take the final exam had on average 0.56 more points in the final exam compared with non-participants. Yet we believe that this self-selection does not constitute an issue in this experiment for several reasons. First, students did not have any information about the treatments or even the existence of the experiment when they had to decide if to answer the questionnaire or not. Thus, students selected themselves into answering a questionnaire and not into receiving monetary incentives for their attainment. Secondly, the selection of the students occurred before the assignment of the different treatments, so it would affect in the same way students independently on their treatment status. Finally, we expect students that do not answer a quick and easy questionnaire in return of 0.2 points in the final exam to be less interested in the course and, consequently, less sensitive to incentives.

The students that hand in the questionnaire are randomly assigned to one of the three treatment groups or the control group. They are informed by email of their assignment status (Figure 1.A3, Figure 1.A4 and Figure 1.A5). In this email, it is explained that the university is going to incentivize the work of the students in that specific course and the type of incentive that they are subject to; however, the existence of the different economic incentives is not divulgated. Because of some technical problems, some students could not receive the email with the treatment status, which leaves us with 1,079[13] participants.

Table 1.1 summarizes the descriptive statistics of the students participating in the experiment. As expected, students attending distance-learning universities are different from the students attending other universities in Spain. They are

[13] 1,104 students successfully received the email with the treatment status. However, 25 of these students did not complete the whole questionnaire, so we are missing some information from them. We have checked that the attrition is balanced among the treatments and control groups. Thus, in order to homogenize our sample sizes for all the outcomes that we will be looking at, we excluded these 25 students from the analysis.

rather older, with a mean age of 33 years. Also, the majority of these students (71.3%) work while they are studying, but only half of them work full-time. Among the students that do not work, 20.9% of them would be willing to work, so unemployment could be one of the main motives for studying.

Table 1.1: SUMMARY STATISTICS

Variable	Obs.	Mean	St. Dev.	Min	Max
Male	1,079	0.453	0.498	0	1
Age	1,079	33.00	8.615	18	62
Age enter UNED	1,079	30.08	7.834	17	61
Grade satisfaction	1,079	7.722	1.232	5	10
Expected grade	1,079	7.074	1.288	3	10
Work	1,079	0.713	0.453	0	1
Willing to work	1,079	0.209	0.406	0	1
Married	1,079	0.282	0.450	0	1
Divorced	1,079	0.0389	0.194	0	1
Single	1,079	0.602	0.490	0	1
Have children	1,079	0.280	0.449	0	1
Finance studies by scholarship	1,079	0.103	0.304	0	1
Finance studies by salary	1,079	0.633	0.482	0	1
Finance studies by family support	1,079	0.154	0.361	0	1
Finance studies by savings	1,079	0.146	0.353	0	1
Low Education Father	1,079	0.813	0.390	0	1
Low Education Mother	1,079	0.870	0.336	0	1
# courses before	1,079	5.267	3.714	0	10
Previous university studies	1,079	0.388	0.488	0	1
Mean previous grades	779	6.599	0.972	5	9.625
Predicted grade	1,079	5.737	0.999	2.414	8.390

Notes: The table reports the summary statistics for all the variables used throughout the empirical analysis. Some students financed their studies by several meaning. Low education of the parents is equal to one if the parents have no universitary studies, and zero otherwise.

Yet, students that take part of this field experiment are not that different from college students that have been incentivized in the previous literature. For instance, 77.7%, 53% and 80% of the students were currently employed while taking part of incentive programs in Canada (Angrist et al., 2009), New York City (Barrow and Rouse, 2013), and Amsterdam (Leuven et al., 2010). Also, in Barrow and Rouse (2013), students were on average 27 years old and 20% of them were married. Moreover, as the main motivation of this paper was to compare different incentive schemes in the same educational setting, the internal validity of the paper is maintained without discussion. Finally, 15.2% of total university students in Spain attend a distance learning university, which makes this pool of students interesting to study by itself.

There are three treatment groups and a control group. In the first treatment group, called "Threshold", 268 students receive an email announcing that all those whose final grade exceeds an 8 (out of 10) will enter a lottery in which they could earn 1,000€. Students also receive information about how many other subjects are in their treatment group. Figure 1.A3 reproduces the email received by the students under this treatment.

In the second treatment group, "Top percentile", 267 subjects receive an email announcing that all students who earn one of the top 25% final grades in their group (one of the 40 best grades in the academic year 2014-15 and one of the 35 best grades in the academic year 2015-16) enter a lottery in which they could earn 1,000€. The email with this treatment is presented in Figure 1.A4. As in all treatments, students also receive information about the number of students they are actually competing with.

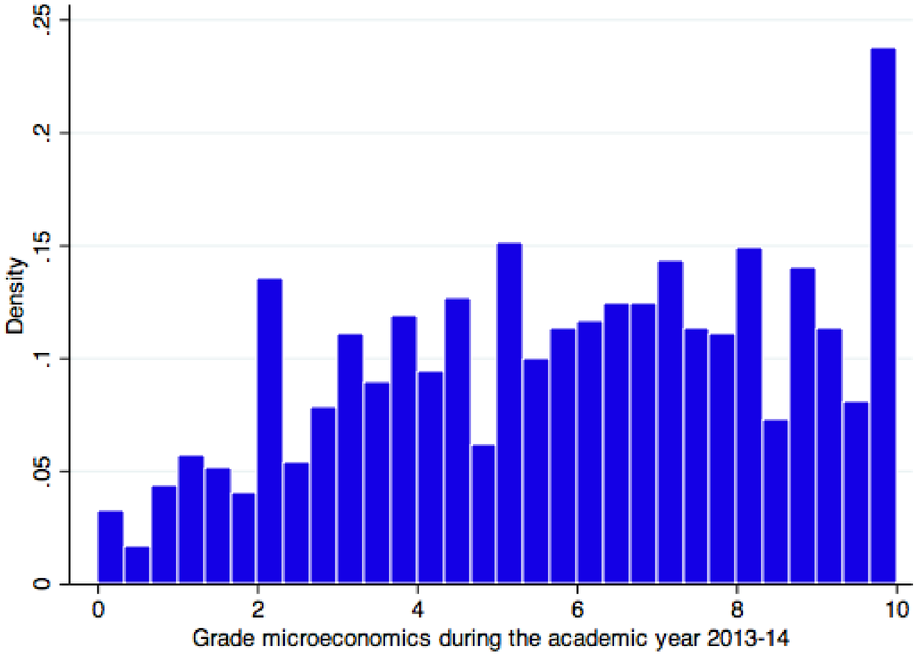
The grade that we select as the threshold in the "Threshold" treatment, 8, is determined so that the expected number of students that would enter the lottery are ex- ante equal to the second treatment group.[14] More specifically, we took the distribution of grades from the same course in the academic year of 2013 (Figure 1.1) as a proxy of the distribution of ability of the students, and the threshold is chosen to be equal to the grade of the 25 top percentile.

In the third treatment group, "Improvement", 277 students receive an email stating that all students who increase their expected grade (indicated in the

[14] 64 students (out of 268) achieved the grade threshold under the "Threshold" treatment while 75 (out of 267) entered in the lottery in the second treatment group.

questionnaire before the experiment was announced) by 1 point will enter a lottery in which they could earn 1,000€. The email with this treatment status can be seen in Figure 1.A4. The improvement of one point was chosen because when using the distribution of grades from 2013, we expect that 25% of the students would achieve this goal so that the number of students entering the lottery is the same across all three treatments. [15]

Figure 1.1: Grades Distribution in Microeconomics during 2013-14



Notes: The graph reports the distribution of grades of the microeconomics course for the academic year 2013-2014.

In the control group, 267 students receive an email with a reminder of the dates of the final exam (Figure 1.A5). This reminder is also present in the other treatments, but we wanted to insure that treatment effects are due to the monetary incentives and are not driven by reading an email.

[15] 62 students (out of 277) increased by one point their expected grade under "Improvement" treatment while 75 (out of 267) entered in the lottery in the second treatment group.

To minimize students' suspicion about the existence of an experiment, all the emails that the students received were signed by a professor of UNED, Marc Vorsatz. He has access to all the data of the course and supervised the whole experiment but was not taking part in the actual teaching process. Also, students were informed that the lotteries will be filmed and sent to them, and that the prize will be paid immediately afterwards via a bank transfer.

The reward is the same for all the three treatment groups. The literature proposes several ways to incentivize college students. Leuven et al. (2010) give the equivalent of 4 (in the low reward condition) to 11€ (in the high reward condition) for each credit students need to pass. Angrist et al. (2014), on the other hand, offer $100 for having a 70 % score in each class. A similar amount is offered by Barrow et al. (2014). De Paola et al. (2012) pay 5€ or 15€ per credit, depending on the treatment. Other studies gave an economic incentive equivalent to the tuition expense for the course (Angrist et al., 2009; Scott-Clayton, 2011). The UNED charges students 12€ per credit of each course. As the course has 6 credits, this means that the total expense of that course for the student is around 72 €.

Given the large number of participants, it is not possible to pay each student a fixed amount if incentives are met. The natural alternative is a lottery. Lottery-based incentives are very common in health promotion plans[16], fundraisers for charities[17], and experiments with large samples of participants. Lotteries have two favorable attributes. First, literature has shown that individuals overestimate low probabilities events[18] (Kahneman and Tversky, 1979). For example, March

[16] For instance, Volpp et al. (2008) offer frequent small prizes, infrequent large prizes or deposit contracts to promote short-term weight loss and find that all payment schemes are equally effective. On the other hand, Niza et al. (2014) find that offering participants a 5 pound voucher is much more likely to result in the screening of a sexually transmitted infection than offering a lottery with a 200 pound prize.

[17] While Landry et al. (2006) show that lotteries raised more money than the voluntary contributions in a door-to-door fund-raising field experiment, Onderstal et al. (2013) finds no significant differences between these two fundraising mechanisms.

[18] We are assuming students are risk neutral. If students were risk averse, fixed rate incentives would be relatively more effective. Conversely, students may be risk loving in the domain of finan-

et al. (2014) find that lotteries with a large prize but small payment likelihood incentivize more than lotteries with the same expected value but more moderate prizes and payment likelihood. Also, Baltussen et al. (2012) shows that this type of payment scheme reduces participants risk-aversion. However, the only paper that look at the differences between fixed rate and lotteries in an educational context, finds no statistical differences between the payment methods (Levitt et al., 2016a). Thus, employing a lottery should not be expected to induce less effort. Secondly, the lottery allows us to fix the budget equally for each incentive scheme, making the comparison easier. Since the budget for each treatment group is 1,000€ per year[19], the ex-ante expected value of the lottery for a student that achieves the grade objective is, more or less, of 27€. As the course that is incentivized has 6 credits, the expected value of the incentive is of 4.5€ per credit, a similar amount to the low reward conditions of Leuven et al. (2010) and De Paola et al. (2012).

The incentivized course started at the beginning of October. From the beginning of November to the beginning of December students had time to complete the questionnaire (Figure 1.A1 and Figure 1.A2). We use the information collected in the questionnaire and the grades of the students in previous courses at UNED (available from their official academic records) to perform the randomization. To check that the randomization was done correctly, we individually regress each variable used for the randomization with respect to the three treatment dummies. Table 1.2 shows that the randomization was successful as we were able to create comparable treatments and control groups with regard to the observable students' characteristics. Nevertheless, for the rest of the analysis, we will be controlling for all these covariates, except for the mean grade in previous courses.[20]

cial rewards (Guryan and Kearney, 2008) in which case the lottery would be even more effective.

[19] Each year the total cost of the experiment was 3,000€.

[20] Note that 23% of the students participating in the experiment did not take any course in UNED before. For these students we do not have information about their previous grades. In the appendix, we show that the randomization was also successful when we only consider those students that took at least some other course in the UNED before. This proves that when we look at the heterogeneous effects by previous grades, the selected group of students is balanced among the treatments and control groups.

Table 1.2: Randomization

Dependent variables	Independent variables			Dependent variables	Independent variables		
	Threshold	Top percentile	Improvement		Threshold	Top percentile	Improvement
Male	0.04	0.02	-0.01	Finance studies by family support	-0.01	0.01	0.00
	(0.04)	(0.04)	(0.04)		(0.03)	(0.03)	(0.03)
Age	-0.28	-0.10	0.15	Finance studies by savings	-0.02	0.03	0.00
	(0.73)	(0.75)	(0.75)		(0.03)	(0.03)	(0.03)
Age enter UNED	-0.58	-0.07	0.13	Low Education Father	0.02	-0.02	0.00
	(0.66)	(0.68)	(0.67)		(0.03)	(0.03)	(0.03)
Work	0.03	-0.01	0.01	Low Education Mother	0.00	-0.06**	-0.02
	(0.04)	(0.04)	(0.04)		(0.03)	(0.03)	(0.03)
Willing to work	0.00	0.03	-0.01	Previous university studies	-0.06	-0.04	-0.04
	(0.03)	(0.04)	(0.03)		(0.04)	(0.04)	(0.04)
Married	-0.01	0.03	0.03	Grade satisfaction	0.05	-0.13	-0.07
	(0.04)	(0.04)	(0.04)		(0.11)	(0.10)	(0.10)
Divorced	-0.02	-0.01	-0.02	Expected grade	-0.01	-0.16	-0.02
	(0.02)	(0.02)	(0.02)		(0.11)	(0.11)	(0.11)
Single	0.02	-0.01	-0.03	Predicted grade	-0.02	-0.08	-0.03
	(0.04)	(0.04)	(0.04)		(0.09)	(0.09)	(0.08)
Have children	0.02	0.01	0.01	Mean previous grades	-0.16	-0.12	-0.11
	(0.04)	(0.04)	(0.04)		(0.10)	(0.10)	(0.10)
Finance studies by scholarship	0.01	0.01	-0.03	# courses before	-0.15	0.00	-0.35
	(0.03)	(0.03)	(0.02)		(0.32)	(0.32)	(0.32)
Finance studies by salary	0.03	-0.06	0.01				
	(0.04)	(0.04)	(0.04)				

Notes: These tables report the results from each regression of the different dependent variable with respect to the three treatment dummies. Each column reports the estimated coefficient from each of the corresponding treatment dummies. * significant at 10%; ** significant at 5%; *** significant at 1%.

We sent the treatment emails on the 11th of December of 2014 and 9th of December of 2015. Students could take a midterm exam at the beginning of December, which is voluntary and can increase their final grade by up to one point. Also, they have continuous contact with the professor via an online forum, where they can ask any type of doubt they have about the course (including content). All the students' interventions could not only been seen and answered by the professor, but also by the rest of the students in the course. Students received two reminders with the treatment conditions. The first one is sent one week before the first exam session, on the 19th of January of 2014 and 18th of January of 2015 and the second one, one week before the second exam session, on the 2th of February of 2014 and 4th of February of 2015.

Then, we have available all the students' information collected in the questionnaires (age, gender, the age at which they entered UNED, their working status, the civil status, the education of their parents, how they finance their studies, the previous university studies, their goal and expected grades). From their official academic records we know the number of course done in UNED before and their grades, if they participate in the voluntary midterm exam and their grade, the exam period they attended, their final exam grade of the incentivized course, and the number of correct, incorrect and answered questions in the final exam. We also have information about the other courses students take during the same term. Finally, we have access to the official forum of the course and we are able to collect the number of interventions of each student.

1.3 Results

1.3.1 General Results

We estimate the average effect of the three types of economic incentives on different performance measures: the probability of taking the final exam, the probability that students take the voluntary midterm, the exam period, the exam grade, and the total number of correct and answered questions in the final exam. We also examine if the incentives influence the number of courses (that are not incentivized) the students take during the same term. Finally, we are able to record the number of interventions in the forum of the incentivized course. This could be considered a proxy of effort during the course, although only 15% of the students participate at least once in the forum.

We use the following econometric model:

$$Y_i = \beta_0 + \beta_2(T1) + \beta_3(T2) + \beta_4(T3) + \alpha X_i + u_i$$

where Y_i is the measure of performance of student i , $T1$, $T2$ and $T3$ are dummy variables that take value one if student i has been assigned to the "Threshold", "Top percentile" or "Improvement" treatment groups or zero otherwise, X_i is a vector of the individual characteristics available, and u_i is an error term. The covariates that are included in the analysis are: age, gender, the age at which students entered UNED, working status, the civil status, the education of the parents, how the students are financed, the previous university studies, the objective and expected grade, the number of courses done in UNED before, and the academic year of the experiment.[21]

We can observe from Table 1.3 that there are almost no average effects for any of the treatments. Students assigned to the "Improvement" seem to answer more questions in the final exam, compared with students in the control group. Also, students under the "Top percentile" incentive scheme do, on average, 0.33 interventions less compared with students that are not incentivized. For the "Threshold" treatment group, the incentive does not impact significantly any of the analyzed performance measures. This result is not surprising because, as pointed out at the beginning of the paper, the majority of the previous literature also does not find average effects.

[21] Table A3 in the Appendix, shows the impact of the three incentives over all the performance outcomes without controlling for the covariates. Controlling for covariates normally increase the power of the still consistent estimates. However, previous literature (Freedman (2008), for instance) points out the importance of comparing the point estimates with and without controls as the controls would induce a finite-sample bias if the treatment effects are heterogeneous. Although the bias is more important for experiments involving fewer than 500 observations. We can observe that the point estimates in the regressions with and without controls are not significantly different.

Table 1.3: AVERAGE EFFECTS

	Take		Exam	Exam	# questions	# questions	# correct by	# courses non	# interventions
	exam	Midterm	Period	Grade	correct	answered	# answered	incentivized	forum
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Threshold	-0.02	-0.02	-0.04	-0.04	-0.03	0.20	-0.01	-0.16	-0.27
	(0.04)	(0.05)	(0.05)	(0.22)	(0.37)	(0.27)	(0.02)	(0.14)	(0.25)
Top percentile	-0.03	0.01	-0.01	-0.07	-0.06	0.04	0.00	-0.05	-0.43*
	(0.04)	(0.05)	(0.05)	(0.20)	(0.34)	(0.27)	(0.02)	(0.14)	(0.23)
Improvement	-0.06	0.01	-0.02	-0.02	0.14	0.59**	-0.02	-0.14	-0.21
	(0.04)	(0.05)	(0.05)	(0.22)	(0.38)	(0.27)	(0.02)	(0.14)	(0.26)
Observations	1,079	1,079	801	801	801	801	801	1,079	1,079
R^2	0.09	0.08	0.03	0.26	0.28	0.15	0.20	0.08	0.04
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES
Mean dep. variable	0.775	0.487	0.367	5.839	12.61	15.99	0.782	1.401	0.674
Std. dev. dep. variable	0.418	0.501	0.483	2.388	4.094	2.952	0.188	1.435	2.626

Notes: The dependent variables are (1) a dummy that takes value 1 if the student has taken the final exam, and 0 otherwise, (2) a dummy that takes value 1 if the student has done the voluntary midterm, and 0 otherwise, (3) a dummy that takes value 1 if the student attended the second period of exams, and 0 otherwise (4) students' exam grade, (5) the number of questions the student answered correctly in the exam, (6) the total number of answered questions in the exam, (7) the number of correct questions divided by the number of answered questions in the exam, (8) the number of courses (non-incentivized) the student has taken during the same term, and (9) total number of interventions in the forum (after the announcement of the treatments) . "Threshold" is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. "Top percentile" is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. "Improvement" is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, the number of courses students have done before in UNED, and the year the experiment took place. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

Why are there no average effects?

There are several reasons that could lead us to not detect a significant average effect of the economic incentives over students' performance.

Firstly, the incentive might not be big enough to compensate the cost of increasing students' effort. The ex-ante expected value of the lottery for a student that achieves the grade objective in each treatment is, more or less, of 27€. As the course that is incentivized has 6 credits, the expected value of the incentive is of 4.5€ per credit, a similar amount to the low reward conditions of Leuven et al. (2010) and De Paola et al. (2012). And De Paola et al. (2012) find positive and significant effects even for their low reward condition. In addition, we expect the lottery to increase the salience of the incentive due to an overestimation of the low probability of winning by the students.[22] Finally, we could argue that if the amount of the incentive mattered to find average effects, we will expect that those students who are more cash constrained (those that do not work but are willing to work) to be more incentivized. However, we find that the average effects do not differ by the working status of the students.[23]

Secondly, there could be effects that we are not able to detect due to the lack of power. Some post-experimental power calculations inform us that we should be able to detect an effect larger or equal to 0.17 standard deviations. This is equivalent to 0.4 points in the final grade or 7 percentage points in the probability of taking the final exam. Very few studies in the previous literature has detected a smaller effect,[24] which could be considered economically small. Thus, we can reject that we are not able to detect medium to large effects of the economic incentives.

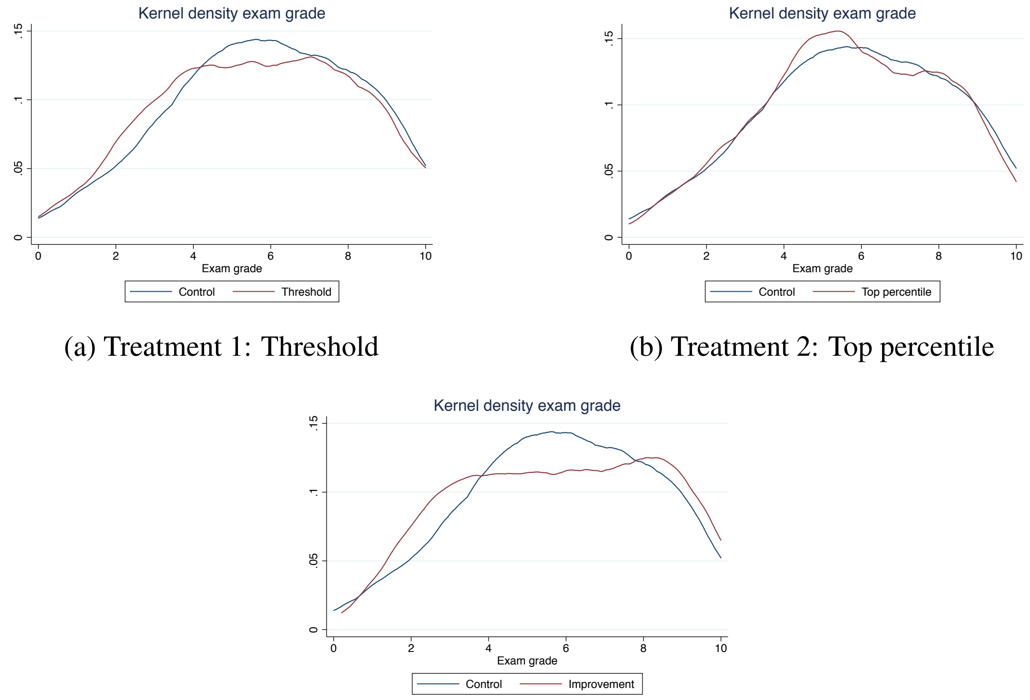
[22] This would be the case under the assumption that students are risk neutral or risk loving. Previous literature has shown that individuals can be risk-loving in the domain of financial rewards (Guryan and Kearney, 2008) in which case the lottery would be relatively more effective.

[23] Results of the average effects by the willingness to work of the student in the Appendix (Table 1.A3). We also show that, even though we did not stratified the randomization by the working status of the student, balanced of the covariates still hold for the different subgroups by their willing to work.

[24] As far as we are aware, only three studies have estimated a smaller effect of the incentives over students achievement. De Paola et al. (2012), Bettinger (2012), and Levitt et al. (2016b) detected an effect of 0.16, 0.15 and 0.1 standard deviations, respectively.

Finally, we might not be capturing any positive (or negative) average effect because the economic incentives might be improving the performance of some type of students while worsen others. In fact, Figure 1.2 shows a plot of the estimation of the kernel distribution of the exam grades for each treatment group compared with the control group. We can perceive some changes in the distribution of the exam grades between the control group and the treatment groups. Thus, in next section, we analyze the heterogeneous effects of the three economic incentives.

Figure 1.2: DENSITY OF EXAM GRADES BY TREATMENT



(a) Treatment 1: Threshold

(b) Treatment 2: Top percentile

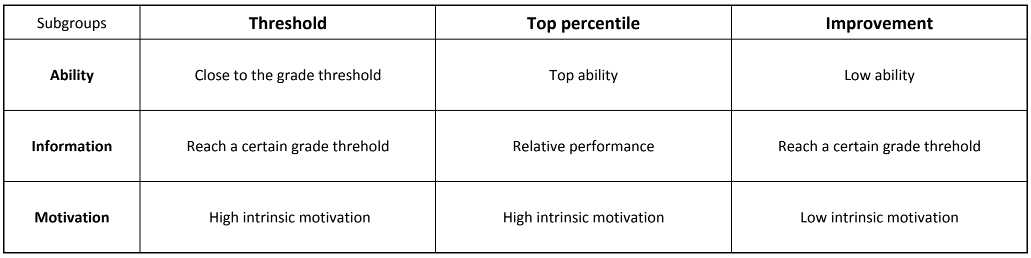
(c) Treatment 3: Improvement

Notes: These figures plot the univariate kernel densities of exam grades for the students in the control group and students under the (a) "Threshold" treatment, (b) "Top percentile" treatment, and (c) "Improvement" treatment.

1.3.2 Heterogeneous Effects

One of the objectives of this paper is to investigate how the different performance targets of the incentive programs interact with the characteristics of the students being incentivized. There are several characteristics of students that previous literature has already pointed out as determinant for incentives to work: intrinsic motivation, experience with the incentivized task, and students' ability. More importantly, we expect these characteristics to have different impacts (predictions summarized in Figure 1.3). Then, in this section, we examine the heterogeneous effects of the economic incentives[25] over intrinsic motivation, experience with the incentivized task, and students' previous performance.

Figure 1.3: PREDICTIONS



Subgroups	**Threshold**	**Top percentile**	**Improvement**
Ability	Close to the grade threshold	Top ability	Low ability
Information	Reach a certain grade threhold	Relative performance	Reach a certain grade threhold
Motivation	High intrinsic motivation	High intrinsic motivation	Low intrinsic motivation

Notes: This table indicates which subgroup of students (in terms of their ability, information, and motivation) we predict to be greater incentivized by each treatment.

Heterogeneous Effects by a Proxy of "Intrinsic Motivation"

Monetary incentives can have two effects on performance. The first one, the direct price effect, makes the incentivized behavior more attractive. The monetary incentive reduces the cost of studying so the student that is paid to study will exert a higher effort compared to the student that is not paid. The second one, the

[25] As covariates we are going to use for the heterogeneous effects are not randomly assigned, they could be correlated with other omitted variables. This is a common problem when looking at heterogeneous effects and should be taken into account when interpreting the interactions. Yet, to minimize the correlation of our proxies of ability, motivation and experience with other characteristics of the students, in all our regressions we will control for all the students' variables we have available.

indirect psychological effect, prays that economic incentives reduce intrinsic motivation (doing something because it is inherently interesting or enjoyable), so the student that is paid is expected to perform worse than the student who is not paid.[26]

Some educators have pointed out the danger of incentivizing students with money, as extrinsic incentives may crowd out intrinsic motivations that are important to producing the desired behavior in this specific context. Many papers mention this possible crowding out effect. However, as far as we know, Bettinger (2012), Fryer (2011), Leuven et al. (2010) and Le (2015) are the only ones that examine it directly, and, with exception of Leuven et al. (2010), the rest find no evidence for the crowding out of intrinsic motivation.

Leuven et al. (2010) borrowed the theoretical framework proposed by Camerer and Hogarth (1999) and suggest that whether the direct price effect is sufficiently large to compensate the loss of intrinsic motivation will depend on the size of the gap between the unincentivized and the incentivized achievement. In our setting, the intrinsic motivation is proxied by the minimal grade students would be satisfied with. We consider this a measure of intrinsic motivation as we expect a student that only feels satisfied with a 9 to be ex ante more intrinsically motivated than a student that feels satisfied with a 5.

Then, we can assume that if the satisfaction grade is close to the incentivized grade, the price effect will be higher than the loss of intrinsic motivation, while if the gap between the satisfaction grade and the incentivized grade is large, the direct price effect will be zero and so the potential loss of intrinsic motivation dominates. As a consequence, we expect high-motivated students under the first and second incentive schemes (“Threshold” and “Top percentile”) to perform better than students with the same motivation in the control group. On the

[26] Previous literature in psychology has widely discuss whether extrinsic rewards crowd out intrinsic motivation in many different contexts. For more information on the different views on the subject see, among others, Deci (1972), Deci and Ryan (1975), Kohn (1999), Gneezy and Rustichini (2000b), or Cameron and Pierce (1994). Some examples on reduced intrinsic motivation in this context could be that the incentive is signaling that the goal is difficult or the student is not able to reach it, or it reduces the joy of studying that the students had before.

other hand, students under the "Improvement" incentive with low motivation and, consequently, low grade expectations, should do better in the final exam compared with similar students in the control group. These predictions are summarized in Figure 1.3.

Table 1.4 shows the heterogeneous effects of the three economic incentives by grade satisfaction. We can observe that for the "Top percentile" treatment, the effect of the economic incentive is positive for the students with a high and negative for the students with a low intrinsic motivation. That is, the effects of the economic incentives go in the same direction as the intrinsic motivation. Therefore, the economic incentives are not crowding out the intrinsic motivation but, when competition takes places, it reinforces it. There are no significant results in the other two treatments. Figure 1.4 shows the graphical representation of these effects.

As a robustness check, we also show in Table 1.A6 the heterogeneous effects dividing the number of previous courses by the median and in terciles. We confirm that the relationship between the incentive schemes and intrinsic motivation is linear. Moreover, we show that the same result is obtained when we perform the same heterogeneous analysis separately for the two different years the experiment took place.[27] Finally, in Table 1.A12 we add to our specification the interactions of all the treatments with predicted grade and the number of courses students passed. We still observe the same positive and significant interaction between motivation and the "Top percentile" incentive. This robustness check proves that our proxy of motivation is capturing a different characteristic of the student that is not explained by our proxies of ability and experience.[28]

[27] To avoid data mining for differential treatment effects, we look for consistent evidence of the heterogeneous treatment effects in the two academic years the experiment took place.

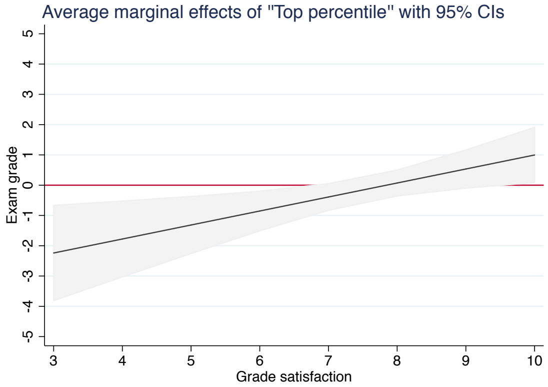
[28] Note that even though our proxies of ability, motivation and experience are not completely orthogonal, the correlation between them is quite low. For instance, the correlation between grade satisfaction and the number of courses passed by the students is 0.03 and 0.23 with predicted grade.

Table 1.4: HETEROGENEOUS EFFECTS BY INTRINSIC MOTIVATION

	Take exam	Exam grade
	(1)	(2)
Threshold	0.09	-2.28*
	(0.23)	(1.36)
Top percentile	0.01	-3.33***
	(0.23)	(1.24)
Improvement	-0.24	-0.67
	(0.24)	(1.54)
Threshold * Grade satisfaction	-0.01	0.29
	(0.03)	(0.18)
Top percentile * Grade satisfaction	-0.00	0.42***
	(0.03)	(0.16)
Improvement * Grade satisfaction	0.02	0.08
	(0.03)	(0.20)
Grade satisfaction	-0.01	-0.04
	(0.02)	(0.15)
Observations	1,079	801
R^2	0.09	0.27
Controls	YES	YES
Mean dep. variable	0.775	5.839
Std. dev. dep. variable	0.418	2.388

Notes: The dependent variables are (1) a dummy that takes value 1 if the student has taken the final exam, and 0 otherwise, and (2) the exam grade. “Grade satisfaction” is the grade that the student will feel satisfied obtaining in the course. “Threshold” is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. “Top percentile” is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. “Improvement” is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, the number of courses students have done before in UNED, and the year the experiment took place. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

Figure 1.4: GRAPHICAL REPRESENTATION OF THE HETEROGENEOUS EFFECTS BY INTRINSIC MOTIVATION



"Top percentile": Exam grade

Notes: Graphical representation of the estimated "Top percentile" treatment effect over the final exam grade for the 8 values of grade satisfaction (from 3 to 10) and its 95% confidence interval.

Further research should investigate why ex-ante high intrinsic motivation enhances performance under the "Top percentile" incentive, while it does not play a role in the other incentive schemes. There are several hypotheses that could explain this differential effect. First, the performance goal under the "Top percentile" scheme could be perceived differently from the other two performance goals. Achieving a grade threshold can be perceived to be less challenging and less prestigious than being in the top of your class. Highly motivated students could be more attracted while low motivated students could be more discouraged by this latter target. A second hypothesis is that the "Top percentile" treatment offers in addition to a monetary reward a piece of previous unknown information about their relative performance. Highly motivated students under the "Top percentile" treatment do not only value the possibility of gaining the prize (also experienced in the "Threshold" condition) but also get a positive utility from knowing that they are among the top students in the course. While low ex-ante motivated students, under the incentive "Top percentile" experience fear of, not only not winning the economic reward, but also the fact that they will know with certainty that they are not the best in the course. Previous literature has already pointed out that providing relative performance feedback enhances performance, even if individuals are not rewarded by their performance (Tran and Zeckhauser, 2012; Azmat and Iriberri, 2010) or they are rewarded using a piece-rate incentive scheme (Azmat and Iriberri, 2016) . A possible follow-up research question could be if information about the possibility of being ranked in the future might induce an increase in performance.

Heterogeneous Effects by a Proxy of "Experience"

Under each incentive scheme, students face different uncertainties when choosing their optimal effort. For instance, in all treatments (including the control group), students might not know their real ability, or their cost function of effort. They have a prior of their ability that could be updated with some signals received during the number of courses previously taken in the same university. Thus, those students with higher experience in UNED might have a more precise belief about their ability.

Still, the type of uncertainty faced by students varies among the incentive schemes. Students under "Top percentile" treatment need to know the distribution of ability in their group to correctly select their optimal effort. Students in the other two treatments do not need information about the distribution of ability in the class to find their optimal effort. On the other hand, students in the "Threshold" and "Improvement" treatments need to know how much effort they require to exert in order to achieve a certain grade threshold. Thus, they need to understand the grade ”production function”. In this setting, while the ”production function” uncertainty can be reduced with experience, student will not increase their knowledge about the ability distribution of their group. In this particular educational setting, students have limited knowledge about the grade distribution of their class and their relative performance, as they do not tend to interact between each other and UNED only informs them about the final grade at the end of each course. Therefore, we hypothesize that the ”Threshold” and ”Improvement” treatments should incentivize more those students with more experience in UNED (Figure 1.3).

Table 1.5 shows that students in the "Threshold" and "Improvement" treatments with more experience have a higher probability of doing the final exam compared with students with the same experience in the control group. Having passed an extra course in UNED before the incentivized course increases by 3 (2) percentage points the probability of showing up in the final exam for students in the "Threshold" ("Improvement") treatment compared with students in the control group. On the other hand, there are no significant heterogeneous effects by experience for the "Top percentile" treatment. Figure 1.5 presents the graphical representation of these linear heterogeneous effects.

As a robustness check, Table 1.A7 in the Appendix shows the heterogeneous effects dividing the number of previous courses by the median and in terciles. We confirm the linearity of the heterogeneous effects of the treatments by the number of previous courses done by the students. We also perform the same heterogeneous effect separately for each year of the experiment and find that the effects are very similar in both years. Furthermore, students could have also learned about their grade production function in other universities that are not UNED. Table 1.A8 shows the heterogeneous effects of the incentive schemes by the students' experience in other universities. Confirming our previous results, we observe that for the "Threshold" and "Improvement" treatments students with previous university studies perform better in the final exam, while students with no previous university studies perform worse. On the other hand, there are no heterogeneous effects for the "Top percentile" incentive, confirming the information mechanism behind previous results. Finally, in Table 1.A12 we add to our specification the interactions of all the treatments with predicted grade and motivation and we still observe the same effects. Then our proxy of experience is capturing a students' characteristic that is not explained by ability and motivation.[29]

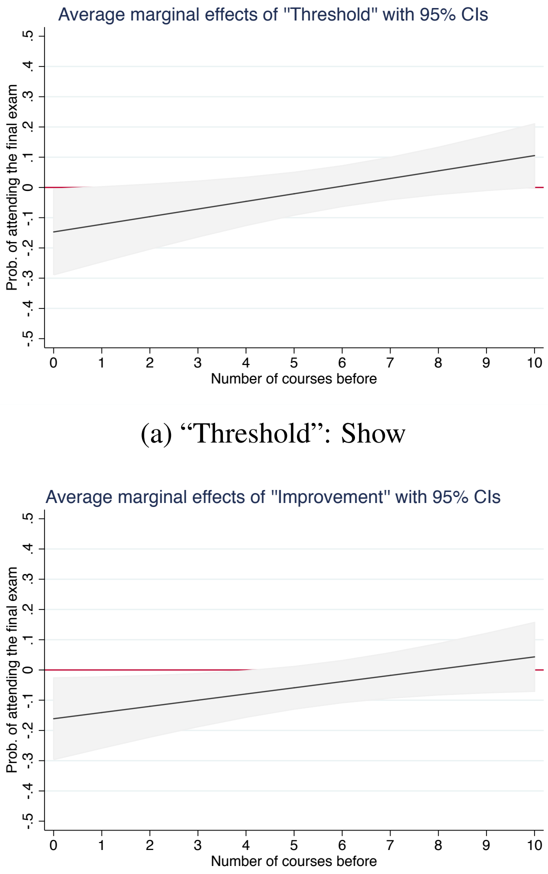
[29] The correlation between the number of courses passed by the students and grade satisfaction is 0.03 and 0.4 with predicted grade.

Table 1.5: HETEROGENEOUS EFFECTS BY EXPERIENCE

	Take exam (1)	Exam grade (2)
Threshold	-0.15**	-0.04
	(0.07)	(0.43)
Top percentile	-0.09	-0.33
	(0.07)	(0.38)
Improvement	-0.17**	0.39
	(0.07)	(0.41)
Threshold * # courses before	0.03**	-0.00
	(0.01)	(0.06)
Top percentile * # courses before	0.01	0.04
	(0.01)	(0.06)
Improvement * # courses before	0.02**	-0.07
	(0.01)	(0.06)
# courses before	0.02***	0.15***
	(0.01)	(0.04)
Observations	1,079	801
R^2	0.10	0.27
Controls	YES	YES
Mean dep. variable	0.775	5.839
Std. dev. dep. variable	0.418	2.388

Notes: The dependent variables are (1) a dummy that takes value 1 if the student has taken the final exam, and 0 otherwise, and (2) the exam grade. "# of courses before" accounts for the number of courses that each student has passed before in UNED. "Threshold" is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. "Top percentile" is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. "Improvement" is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, the number of courses students have done before in UNED, and the year the experiment took place. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

Figure 1.5: GRAPHICAL REPRESENTATION OF THE HETEROGENEOUS EFFECTS BY EXPERIENCE IN UNED



(a) "Threshold": Show

(b) "Improvement": Show

Notes: Graphical representation of the estimated (a) "Threshold", and (b) "Improvement" treatment effect over the probability of taking the final exam for the 11 values of the number of courses previously passed in UNED (from 0 to 10) and its 95% confidence interval.

These results prove that experience with the incentivized task reduces the uncertainty about the grade "production function" but does not helps to better understand the distribution of ability of the students in the class. This is consistent with the fact that students do not interact between each other during any course and their information about their peers is minimal. Previous literature has already pointed out the importance of experience of students for incentives to improve performance. Angrist et al. (2014) performed a randomized control trial with first and second year student in a Canadian university and they found a positive impact of the economic incentive only for second year students. They hypothesize that first year students might have not responded to the program because they have not yet developed successful study techniques which could increase the cost associated with improving their grade or second year students may "have better sense of how to improve their grades" (pag. 29, Angrist et al. (2014)). In the context of teachers incentives, Bellés-Obrero and Lombardi (2016) also argue that teachers' inexperience with the incentivized task could be one of the reasons why the program had no impact.

Heterogeneous Effects by Proxies of "Ability"

Most of the previous literature on economic incentives on education examines heterogeneous effects of the incentives by ability of the student. Some studies find that the students' ability has no effect on individual performance (Angrist et al., 2009; Barrow and Rouse, 2013; Fryer, 2011). Others show that the economic incentives have positive effects for high ability students, while they do not impact low ability students at all (Angrist and Lavy, 2009; Bettinger, 2012; De Paola et al., 2012; Kremer et al., 2009). Finally, Leuven et al. (2010) find that the economic incentive impacts positively high ability and negatively low ability students.

From a theoretical point of view[30] (as summarized in Figure 1.3), treatments incentivize different students. The "Threshold" treatment should only incentivize students whose ability is close to the grade threshold. Student with ability higher than the threshold will get the incentive independently of whether or not they are exerting a higher effort and low ability students will need a very high effort in order to reach the threshold, and the cost of that effort exceeds the benefit obtained by the economic incentive. On the other hand, the "Top percentile" treatment should incentivize all students in the top of the distribution. Unlike the "Threshold" incentive, students under this incentive scheme are uncertain about the grade target they need to reach, as it will depend on the distribution of ability

[30] Predictions are based on the results of an on-going theoretical model.

and effort of all the students in the group. Thus, students with very high ability, driven by the potential competition, will be pushed to exert more effort. Finally, under the assumption of increasing marginal cost of effort, the "Improvement" treatment could induce a higher performance of students with a low ability. Students with less ability will expect ex-ante a lower grade and increasing a point from a lower grade is less costly than increasing a point from a higher grade.

For our proxy of ability, we follow the approach used by Angrist and Lavy (2009), De Paola et al. (2012), Barrow and Rouse (2013) and Levitt et al. (2016a) to estimate the predicted performance of students.[31] We first take only the students in the control group and regress the exam grade against a vector of all individual characteristics that we have available. We estimate the coefficients of that model and then use them to predict the exam grade of all subjects. That is, we are constructing a counterfactual of what would have been the grade of the students in the case that the incentive would have not taken place.

Table 1.6 shows that, for all the treatments, students with higher or lower predicted grade do not significantly perform differently under the economic incentives when compared with the control group. If anything, the interaction treatment effect is negative (although no significant). On the other hand, students with a higher predicted ability answer fewer questions in the final exam than the students with the same predicted ability in the control group. Moreover, students with lower predicted ability answer more questions in all treatment groups than in the control group. We observe that the linear interaction between the three

[31] We could consider the mean grade of all the courses students have done at UNED as a proxy of ability. Yet, this measure has mainly two problems. The first one is that not all the students have done the same number of courses before. UNED allows for a higher flexibility compared to standard universities. Students can take any course in the order they prefer, even second year courses without having to take any first year course. In fact, 24% haven't taken any course in UNED before the incentivized course. For these 257 students I do not have any measure of ability. Moreover, there exists a great dispersion in the number of courses students have taken before, which goes from 0 to 10. The second problem is that we only observe grades for those courses where the student passed the exam. So we have no information if the student has done the course but didn't pass. For these two reasons we think this proxy might not be reflecting correctly the inherent ability of the students. Anyway, Table 1.A9 shows the linear interaction between the treatments and the mean grade of previous courses and there are no differential treatments effects.

treatment dummies and the predicted grade is highly significant. Finally, since the predicted grade has no significant effect on the percentage of questions that are answered correctly, it is coherent that there are no effects on the final grade. Figure 1.6 shows the linear graphical representation of these heterogeneous effects.

As a robustness check, in the Appendix (Table 1.A10), we also look at the heterogeneous effects dividing the predicted grade by the median or in terciles. We confirm that for the second and third treatments the relationship between the incentives and the mean grade seems to be quite linear. However, we do not observe the same for the ”Threshold” incentive. As this incentive should only incentivize students around the threshold, we only observe that the number of answered question in the final exam is significantly higher for those students with a predicted grade in the third quartile. Finally, in Table 1.A11, we show that this heterogeneous effect is replicable if we look separately at each academic year.

The heterogeneous effects of the total number of answered questions by predicted grade has a straightforward explanation in terms of marginal utility. Unlike the previous literature, in this experiment we are incentivizing performance in a multiple-choice exam with penalties for the incorrect answers. In this type of exams, the decision on how many questions to answer is a strategic decision. Students, when deciding whether to answer a question or not, have two options. They do not answer and get a sure outcome of 0 points. If they answer, it will be correct with probability p and they will gain 0.5 points or it will be incorrect with probability $(1-p)$ and they will lose 0.15 points.

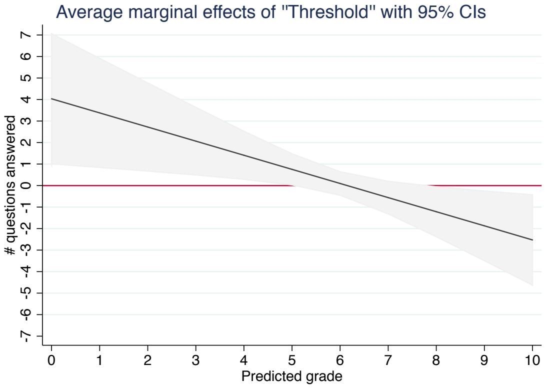
Once economic incentives are introduced, the expected utilities of answering correctly or incorrectly are modified, since students do not only care about gaining or losing points in the final grade, but also the probability of gaining or losing the incentive. High ability students, have ex-ante high probability of winning the economic incentive so their disutility of answering incorrectly is very high. They know that if they risk it, even with a very low probability, they could lose the economic incentive.

Table 1.6: Heterogeneous Effects by Predicted Grade

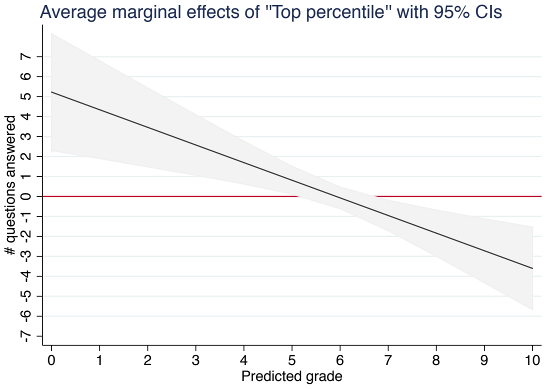
	Exam grade	# questions answered	Perc. questions correct
	(1)	(2)	(3)
Threshold	1.97	4.03***	0.04
	(1.25)	(1.54)	(0.10)
Top percentile	0.89	5.23***	-0.03
	(1.09)	(1.49)	(0.10)
Improvement	1.77	5.98***	-0.02
	(1.19)	(1.45)	(0.10)
Threshold *Predicted grade	-0.36*	-0.66**	-0.01
	(0.21)	(0.26)	(0.02)
Top percentile *Predicted grade	-0.16	-0.88***	0.01
	(0.18)	(0.25)	(0.02)
Improvement *Predicted grade	-0.31	-0.93***	-0.00
	(0.20)	(0.24)	(0.02)
Predicted grade	1.00***	1.19***	0.05***
	(0.12)	(0.16)	(0.01)
Observations	801	801	801
R^2	0.13	0.07	0.10
Controls	NO	NO	NO
Mean dep. variable	5.839	15.99	0.782
Std. dev. dep. variable	2.388	2.952	0.188

Notes: The dependent variables are (1) the exam grade, (2) total number of questions answered in the exam, and (3) the total number of correct questions divided by the total number of questions answered in the exam. "Predicted grade" is the prediction of the grade the students under the incentive schemes would have obtained in the absence of the incentive. This predicted grade is calculated following the methodology used by Angrist and Lavy (2009). "Threshold" is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. "Top percentile" is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. "Improvement" is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

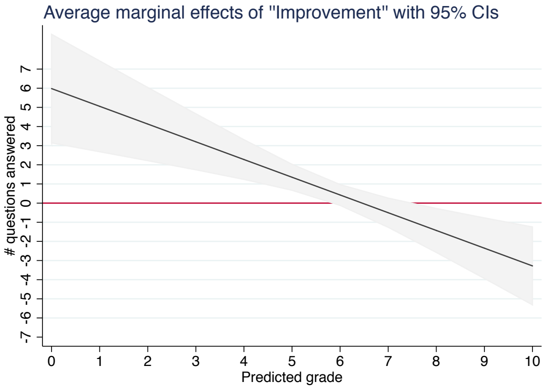
Figure 1.6: Graphical Representation of the Heterogeneous Effects by Predicted Grade



(a) “Threshold”: Questions answered



(b) “Top percentile”: Questions answered



(c) “Improvement”: Questions answered

Notes: Graphical representation of the estimated (a) “Threshold”, (b) “Top percentile”, and (c) “Improvement” treatment effect over the total number of answered questions for the 11 values of the predicted exam grade (from 0 to 10) and its 95% confidence interval.

Then, treated students will answer fewer questions than untreated students. On the other hand, low ability students are very far away from winning the incentive so the disutility of answering incorrectly is very low. However, the expected utility of answering correctly, by chance, is very large as they could win the incentive. These students have a hope of a large gain. Thus, for these students the economic incentive increases the number of answered questions.

This explanation is consistent with prospect theory. Kahneman and Tversky (1979) state that individuals perceive outcomes as losses or gains, instead of final states of welfare. Losses and gains are defined as deviations from a reference point that is normally the current position of the individual. In our setting, incentives change the reference point of students differently depending on their ability. High ability students have as their reference point winning the incentive, while low ability students' reference point is to not win the incentive. Thus, after the incentive, the expected utility of answering an additional question entails a low probability of a very large loss for high ability individuals and a very low probability of a very large gain for low ability students.

To confirm this strategic behavior of students, we order the number of multiple-choice questions of the exam by the level of difficulty. We determine the level of difficulty of a specific question by the percentage of correct answers of students in the control group. We considered the 5 questions with the lowest percentage of correct answers to be the most difficult, and the 5 questions with the highest percentage of correct answers to be the easiest. The remaining 10 questions have a medium level of difficulty. Table 1.7 shows the heterogeneous effects by predicted grade of the total number of easy, medium, and difficult questions answered, as well as, the percentage of easy, medium, and difficult correct questions. Students with a higher predicted ability answer fewer medium and difficult questions in the final exam than the students with the same predicted ability in the control group. These students also have a lower percentage of correct medium and difficult answers. We do not observe the same effect for the easy questions, confirming the strategic behavior of the students.

Table 1.7: HETEROGENEOUS EFFECTS BY PREDICTED GRADE: DIFFICULTY OF QUESTIONS

	# questions answered			Perc. correct answers		
	Easy	Medium	Difficult	Easy	Medium	Difficult
	(1)	(2)	(3)	(4)	(5)	(6)
Threshold	0.02	1.78*	1.95***	0.03	0.17	0.46***
	(0.22)	(1.03)	(0.72)	(0.11)	(0.14)	(0.15)
Top percentile	0.27	3.07***	1.93***	-0.05	0.23*	0.25**
	(0.20)	(0.99)	(0.73)	(0.11)	(0.13)	(0.13)
Improvement	0.24	3.58***	2.29***	-0.08	0.32**	0.33**
	(0.23)	(1.02)	(0.70)	(0.13)	(0.14)	(0.14)
Threshold *Predicted grade	-0.01	-0.30*	-0.31**	-0.01	-0.03	-0.07***
	(0.03)	(0.17)	(0.12)	(0.02)	(0.02)	(0.03)
Top percentile *Predicted grade	-0.04	-0.52***	-0.34***	0.00	-0.04*	-0.04*
	(0.03)	(0.16)	(0.12)	(0.02)	(0.02)	(0.02)
Improvement *Predicted grade	-0.05	-0.57***	-0.33***	0.00	-0.05**	-0.04*
	(0.04)	(0.17)	(0.12)	(0.02)	(0.02)	(0.02)
Predicted grade	0.05**	0.77***	0.38***	0.04***	0.23***	0.11***
	(0.02)	(0.11)	(0.09)	(0.01)	(0.03)	(0.02)
Observations	801	801	801	801	801	801
R^2	0.01	0.07	0.04	0.06	0.13	0.10
Controls	YES	YES	YES	YES	YES	YES
Mean dep. variable	4.841	8.010	3.188	0.880	0.640	0.341
Std. dev. dep. variable	0.405	2.002	1.354	0.205	0.278	0.267

Notes: The dependent variables are (1-3) the total number of questions answered in the exam, and (4-6) the total number of correct questions divided by the total number of questions answered in the exam. The sample is divided by the difficulty of the questions of the exam, measured by the percentage of correct questions of students in the control group. “Predicted grade” is the prediction of the grade the students under the incentive schemes would have obtained in the absence of the incentive. This predicted grade is calculated following the methodology used by Angrist and Lavy (2009). “Threshold” is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. “Top percentile” is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. “Improvement” is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

This result has important implications. Until now part of the literature has argued that economic incentives do not work in the educational environment. It has been mainly argued that students might not know the best way to improve their attainment, or feel that exam grades are determined mainly by factors outside their control (Fryer, 2011). Here we show that strategic behavior induced by the incentives is another possible reason why incentives do not impact students exam grades.

1.4 Conclusion

Despite the large literature on incentives in education, the effect of economic incentives on students' performance is still far from unequivocal (Gneezy et al., 2011). While the previous literature has mainly focused on the question of whether incentive programs improve students' performance, this paper studies how the characteristics of different incentive schemes interact with those of the persons being incentivized.

We conduct a randomized control trial at a distance learning university in Spain to compare three monetary incentives with different performance targets for students. Students need to achieve at least an 8 out of 10 in the exam ("Threshold"), be in the top 25% of their group ("Top percentile"), or increase by 1 point their expected grade ("Improvement") in order to enter a lottery of 1,000€.

We find no average effect for any of the incentives, but there are interactions between the types of students and incentive treatments. We show that for the "Top percentile" treatment, the effect of the economic incentive is positive for students with high intrinsic motivation and negative for students with low intrinsic motivation. In addition, we show that students in the "Threshold" and "Improvement" treatments with more (less) experience with the incentivized task have a higher (lower) probability of doing the final exam compared with students with the same experience in the control group. Interestingly, we show that economic rewards from a multiple choice exam with penalties for incorrect answers change the marginal utility of answering additional questions in the final exam differently for students of diverse ability. High ability students, having a high disutility of answering incorrectly, answer fewer questions with the economic incentive than without it. On the other hand, low ability students are more prepared to gamble on answering questions for which they do not know the correct responses and thus increase the number of questions answered in the final exam. In other words, incentives foster students' strategic behavior that is triggered by the way performance is measured.

This study adds to the vast literature on the effects of monetary incentives on students in several ways. First, the main contribution is the comparison of different incentives schemes with alternative performance targets. Previous studies not only differ on the incentive scheme examined but also in the economic and social context where the experiments were conducted, the duration, or the size of the cash reward. This paper highlights that student performance targets in incentive programs matter, independently of other design features. Particularly, we show that performance incentives interact with characteristics of the students being incentivized, such as intrinsic motivation and experience with the incentivized task. Second, the proliferation of online learning platforms constitutes a great opportunity for experiments with several treatment arms controlling for potential spillovers and at a lower cost. Finally, this paper highlights the need to pay more attention to the task that is incentivized. If the task allows for strategic behavior by students, the incentive might foster this type of behavior instead of learning.

The question of whether monetary incentives are effective in increasing students' attainment is too narrow and further research is needed to identify the features of incentive design that matter in practice as well as how different design features interact. This study highlights the potential benefits for future policy development of understanding the mechanisms through which students respond to these interventions. For instance, since incentives with a competitive component are beneficial for highly motivated students but detrimental for poorly motivated students, any educational policy that introduces competition between students should be accompanied by programs that address the lack of student motivation. Another natural question for further research is how the results of these types of programs compare with other policies that also have the objective of increasing students' attainment. Since these policies target different populations and endpoints, these comparisons will need further experimentation.

Appendix Figures and Tables

Table 1.A1: RANDOMIZATION FOR STUDENTS THAT HAVE GRADES FROM PREVIOUS COURSES IN UNED

Dependent variables	Independent variables			Dependent variables	Independent variables		
	Threshold	Top percentile	Improvement		Threshold	Top percentile	Improvement
Male	0.03 (0.05)	-0.01 (0.05)	-0.01 (0.05)	Finance studies by salary	0.02 (0.05)	-0.09* (0.05)	0.01 (0.05)
Age	-0.48 (0.84)	0.46 (0.88)	0.50 (0.87)	Finance studies by family support	0.01 (0.03)	0.02 (0.03)	0.00 (0.03)
Age enter UNED	-0.74 (0.76)	0.55 (0.83)	0.27 (0.80)	Finance studies by savings	-0.04 (0.03)	0.03 (0.04)	-0.02 (0.04)
Work	0.02 (0.04)	-0.05 (0.05)	0.00 (0.04)	Low Education Father	0.02 (0.04)	-0.02 (0.04)	-0.02 (0.04)
Willing to work	-0.02 (0.04)	0.03 (0.04)	-0.01 (0.04)	Low Education Mother	0.00 (0.03)	-0.03 (0.03)	-0.04 (0.03)
Married	-0.03 (0.05)	0.02 (0.05)	0.03 (0.05)	Grade satisfaction	0.13 (0.13)	0.00 (0.12)	-0.05 (0.12)
Divorced	-0.01 (0.02)	-0.01 (0.02)	-0.02 (0.02)	Expected grade	0.05 (0.13)	0.00 (0.13)	0.03 (0.13)
Single	0.03 (0.05)	0.00 (0.05)	-0.03 (0.05)	Predicted grade	0.01 (0.10)	-0.05 (0.10)	0.00 (0.09)
Have children	0.00 (0.05)	0.00 (0.05)	-0.02 (0.05)	# courses before	-0.04 (0.27)	-0.03 (0.27)	-0.09 (0.27)
Finance studies by scholarship	0.03 (0.03)	0.03 (0.03)	0.00 (0.03)				

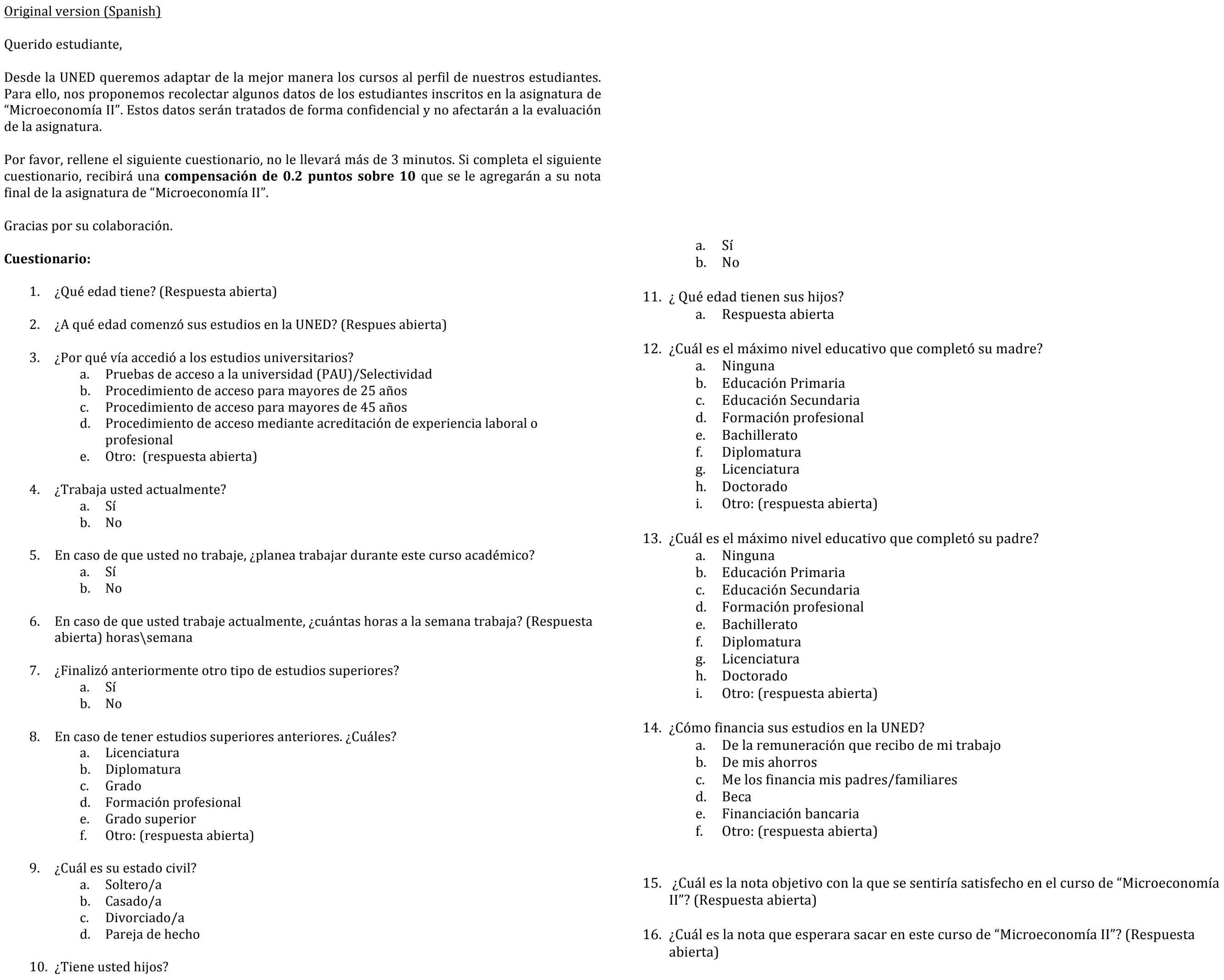
Notes: These tables report the results from each regression of the different dependent variables with respect to the three treatment dummies for only those students that have done at least one previous course in UNED before the incentivized one. Each column reports the estimated coefficient from each of the corresponding treatment dummies.* significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A2: Randomization for Students that Took the Final Exam

Dependent variables	Independent variables			Dependent variables	Independent variables		
	Threshold	Top percentile	Improvement		Threshold	Top percentile	Improvement
Male	0.01	0.03	-0.04	Finance studies by family support	-0.01	0.01	0.00
	(0.05)	(0.05)	(0.05)		(0.03)	(0.03)	(0.03)
Age	0.71	0.20	0.12	Finance studies by savings	-0.02	0.03	0.00
	(0.86)	(0.86)	(0.87)		(0.03)	(0.03)	(0.03)
Age enter UNED	0.01	0.07	0.12	Low Education Father	0.02	-0.02	0.00
	(0.77)	(0.78)	(0.77)		(0.03)	(0.03)	(0.03)
Work	0.04	0.00	0.04	Low Education Mother	0.00	-0.06**	-0.02
	(0.04)	(0.05)	(0.05)		(0.03)	(0.03)	(0.03)
Willing to work	-0.03	0.02	-0.04	Previous university studies	-0.06	-0.04	-0.04
	(0.04)	(0.04)	(0.04)		(0.04)	(0.04)	(0.04)
Married	0.02	0.07	0.03	Grade satisfaction	0.05	-0.13	-0.07
	(0.04)	(0.04)	(0.04)		(0.11)	(0.10)	(0.10)
Divorced	-0.01	-0.02	-0.02	Expected grade	-0.01	-0.16	-0.02
	(0.02)	(0.02)	(0.02)		(0.11)	(0.11)	(0.11)
Single	-0.01	-0.03	-0.03	Predicted grade	-0.02	-0.08	-0.03
	(0.05)	(0.05)	(0.05)		(0.09)	(0.09)	(0.08)
Have children	0.06	0.02	0.01	Mean previous grades	-0.16	-0.12	-0.11
	(0.04)	(0.04)	(0.04)		(0.10)	(0.10)	(0.10)
Finance studies by scholarship	0.01	0.00	-0.02	# courses before	-0.15	0.00	-0.35
	(0.03)	(0.03)	(0.03)		(0.32)	(0.32)	(0.32)
Finance studies by salary	0.03	-0.06	0.01				
	(0.04)	(0.04)	(0.04)				

Notes: These tables report the results from each regression of the different dependent variables with respect to the three treatment dummies for only those students that took the final exam. Each column reports the estimated coefficient from each of the corresponding treatment dummies.* significant at 10%; ** significant at 5%; *** significant at 1%.

Figure 1.A1: QUESTIONNAIRE: SPANISH VERSION



Original version (Spanish)

Querido estudiante,

Desde la UNED queremos adaptar de la mejor manera los cursos al perfil de nuestros estudiantes. Para ello, nos proponemos recolectar algunos datos de los estudiantes inscritos en la asignatura de “Microeconomía II”. Estos datos serán tratados de forma confidencial y no afectarán a la evaluación de la asignatura.

Por favor, rellene el siguiente cuestionario, no le llevará más de 3 minutos. Si completa el siguiente cuestionario, recibirá una **compensación de 0.2 puntos sobre 10** que se le agregarán a su nota final de la asignatura de “Microeconomía II”.

Gracias por su colaboración.

Cuestionario:

1. ¿Qué edad tiene? (Respuesta abierta)

2. ¿A qué edad comenzó sus estudios en la UNED? (Respues abierta)

3. ¿Por qué vía accedió a los estudios universitarios?
   a. Pruebas de acceso a la universidad (PAU)/Selectividad
   b. Procedimiento de acceso para mayores de 25 años
   c. Procedimiento de acceso para mayores de 45 años
   d. Procedimiento de acceso mediante acreditación de experiencia laboral o profesional
   e. Otro: (respuesta abierta)

4. ¿Trabaja usted actualmente?
   a. Sí
   b. No

5. En caso de que usted no trabaje, ¿planea trabajar durante este curso académico?
   a. Sí
   b. No

6. En caso de que usted trabaje actualmente, ¿cuántas horas a la semana trabaja? (Respuesta abierta) horas\semana

7. ¿Finalizó anteriormente otro tipo de estudios superiores?
   a. Sí
   b. No

8. En caso de tener estudios superiores anteriores. ¿Cuáles?
   a. Licenciatura
   b. Diplomatura
   c. Grado
   d. Formación profesional
   e. Grado superior
   f. Otro: (respuesta abierta)

9. ¿Cuál es su estado civil?
   a. Soltero/a
   b. Casado/a
   c. Divorciado/a
   d. Pareja de hecho

10. ¿Tiene usted hijos?
    a. Sí
    b. No

11. ¿ Qué edad tienen sus hijos?
    a. Respuesta abierta

12. ¿Cuál es el máximo nivel educativo que completó su madre?
    a. Ninguna
    b. Educación Primaria
    c. Educación Secundaria
    d. Formación profesional
    e. Bachillerato
    f. Diplomatura
    g. Licenciatura
    h. Doctorado
    i. Otro: (respuesta abierta)

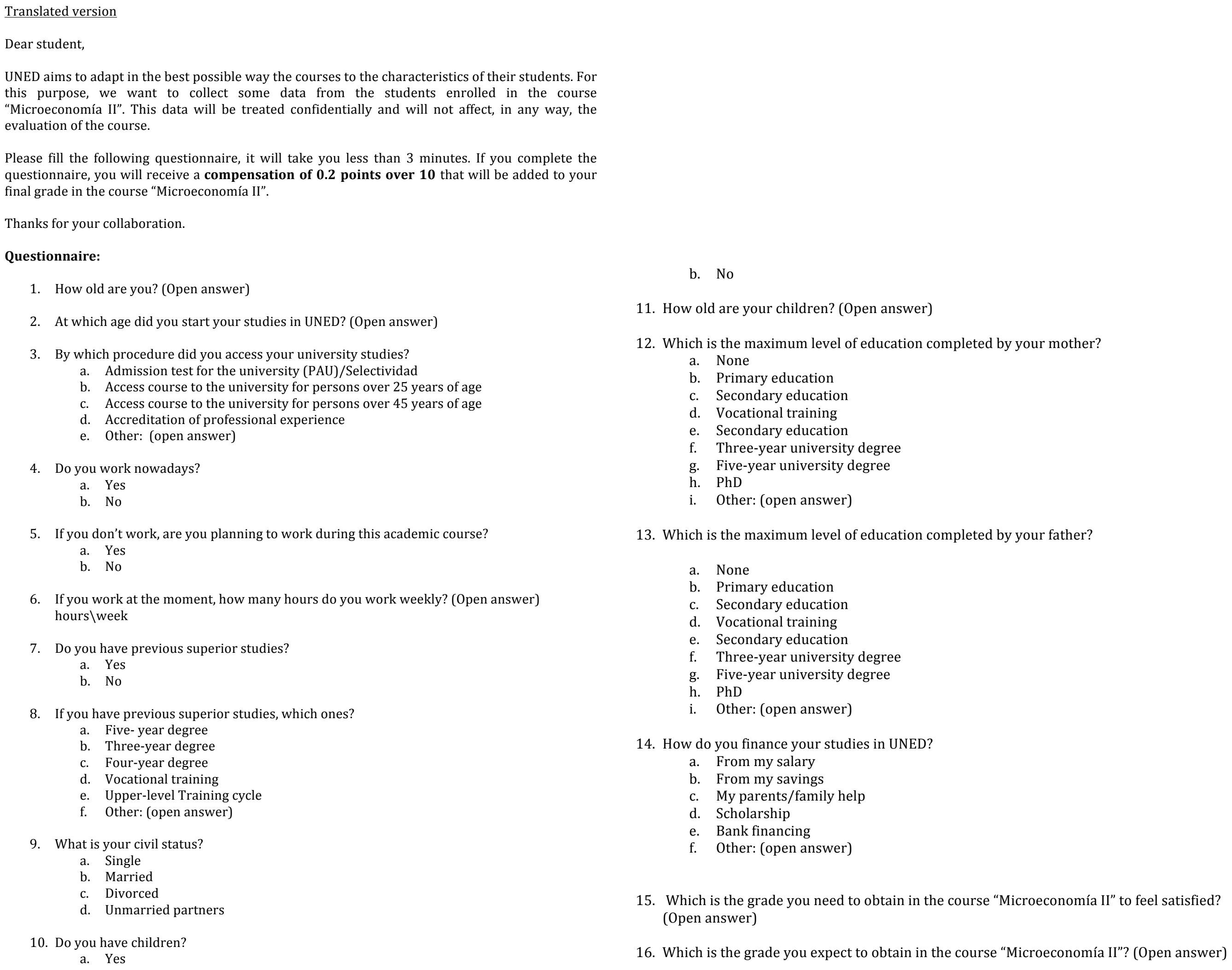
13. ¿Cuál es el máximo nivel educativo que completó su padre?
    a. Ninguna
    b. Educación Primaria
    c. Educación Secundaria
    d. Formación profesional
    e. Bachillerato
    f. Diplomatura
    g. Licenciatura
    h. Doctorado
    i. Otro: (respuesta abierta)

14. ¿Cómo financia sus estudios en la UNED?
    a. De la remuneración que recibo de mi trabajo
    b. De mis ahorros
    c. Me los financia mis padres/familiares
    d. Beca
    e. Financiación bancaria
    f. Otro: (respuesta abierta)

15. ¿Cuál es la nota objetivo con la que se sentiría satisfecho en el curso de “Microeconomía II”? (Respuesta abierta)

16. ¿Cuál es la nota que esperara sacar en este curso de “Microeconomía II”? (Respuesta abierta)

Figure 1.A2: QUESTIONNAIRE: ENGLISH VERSION



Translated version

Dear student,

UNED aims to adapt in the best possible way the courses to the characteristics of their students. For this purpose, we want to collect some data from the students enrolled in the course “Microeconomía II”. This data will be treated confidentially and will not affect, in any way, the evaluation of the course.

Please fill the following questionnaire, it will take you less than 3 minutes. If you complete the questionnaire, you will receive a **compensation of 0.2 points over 10** that will be added to your final grade in the course “Microeconomía II”.

Thanks for your collaboration.

Questionnaire:

1. How old are you? (Open answer)

2. At which age did you start your studies in UNED? (Open answer)

3. By which procedure did you access your university studies?
   a. Admission test for the university (PAU)/Selectividad
   b. Access course to the university for persons over 25 years of age
   c. Access course to the university for persons over 45 years of age
   d. Accreditation of professional experience
   e. Other: (open answer)

4. Do you work nowadays?
   a. Yes
   b. No

5. If you don’t work, are you planning to work during this academic course?
   a. Yes
   b. No

6. If you work at the moment, how many hours do you work weekly? (Open answer) hours\week

7. Do you have previous superior studies?
   a. Yes
   b. No

8. If you have previous superior studies, which ones?
   a. Five- year degree
   b. Three-year degree
   c. Four-year degree
   d. Vocational training
   e. Upper-level Training cycle
   f. Other: (open answer)

9. What is your civil status?
   a. Single
   b. Married
   c. Divorced
   d. Unmarried partners

10. Do you have children?
    a. Yes
    b. No

11. How old are your children? (Open answer)

12. Which is the maximum level of education completed by your mother?
    a. None
    b. Primary education
    c. Secondary education
    d. Vocational training
    e. Secondary education
    f. Three-year university degree
    g. Five-year university degree
    h. PhD
    i. Other: (open answer)

13. Which is the maximum level of education completed by your father?

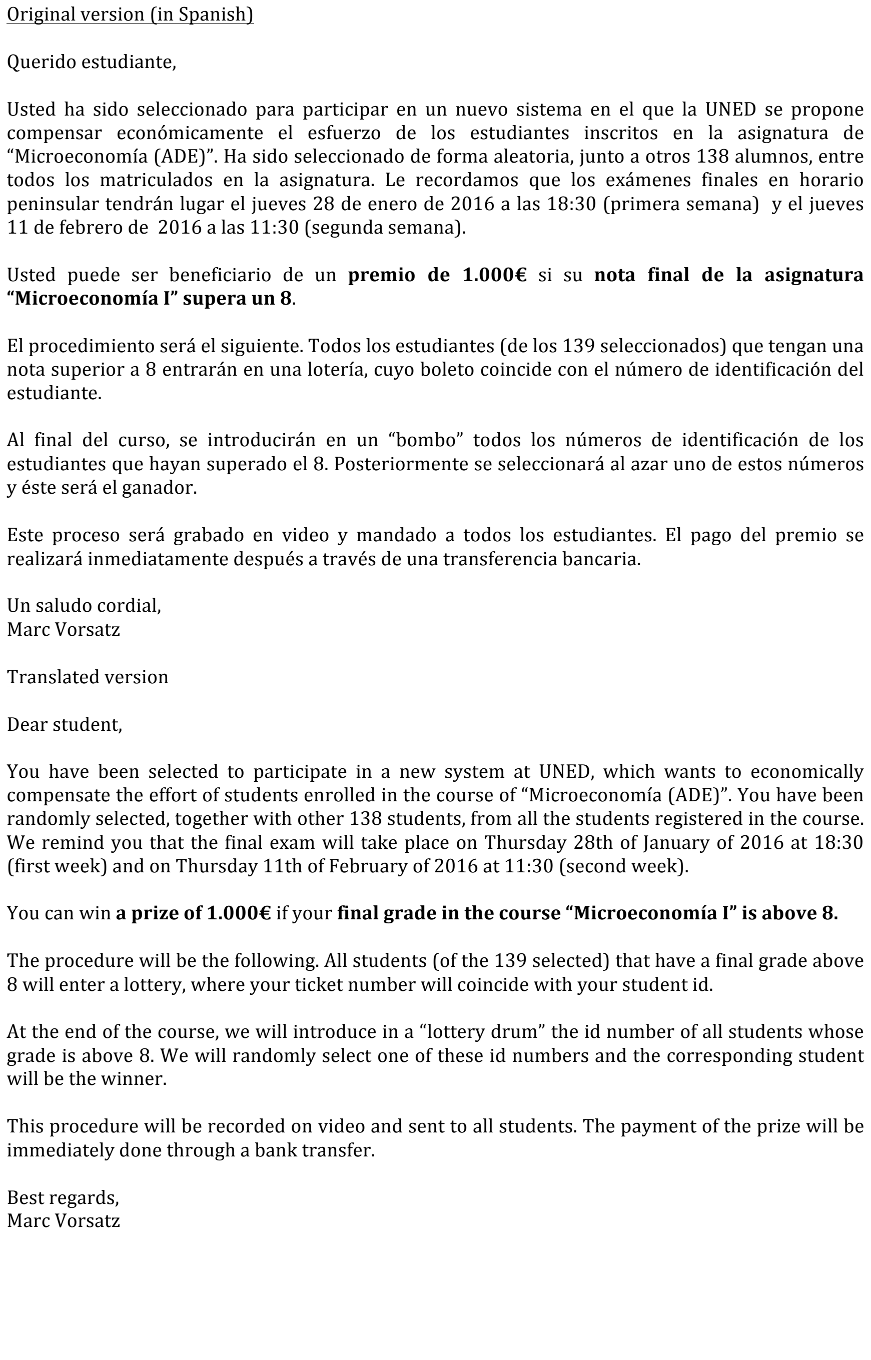
    a. None
    b. Primary education
    c. Secondary education
    d. Vocational training
    e. Secondary education
    f. Three-year university degree
    g. Five-year university degree
    h. PhD
    i. Other: (open answer)

14. How do you finance your studies in UNED?
    a. From my salary
    b. From my savings
    c. My parents/family help
    d. Scholarship
    e. Bank financing
    f. Other: (open answer)

15. Which is the grade you need to obtain in the course “Microeconomía II” to feel satisfied? (Open answer)

16. Which is the grade you expect to obtain in the course “Microeconomía II”? (Open answer)

Figure 1.A3: EMAIL TREATMENT 1: THRESHOLD



Original version (in Spanish)

Querido estudiante,

Usted ha sido seleccionado para participar en un nuevo sistema en el que la UNED se propone compensar económicamente el esfuerzo de los estudiantes inscritos en la asignatura de “Microeconomía (ADE)”. Ha sido seleccionado de forma aleatoria, junto a otros 138 alumnos, entre todos los matriculados en la asignatura. Le recordamos que los exámenes finales en horario peninsular tendrán lugar el jueves 28 de enero de 2016 a las 18:30 (primera semana) y el jueves 11 de febrero de 2016 a las 11:30 (segunda semana).

Usted puede ser beneficiario de un **premio de 1.000€** si su **nota final de la asignatura “Microeconomía I” supera un 8**.

El procedimiento será el siguiente. Todos los estudiantes (de los 139 seleccionados) que tengan una nota superior a 8 entrarán en una lotería, cuyo boleto coincide con el número de identificación del estudiante.

Al final del curso, se introducirán en un “bombo” todos los números de identificación de los estudiantes que hayan superado el 8. Posteriormente se seleccionará al azar uno de estos números y éste será el ganador.

Este proceso será grabado en video y mandado a todos los estudiantes. El pago del premio se realizará inmediatamente después a través de una transferencia bancaria.

Un saludo cordial,
Marc Vorsatz

Translated version

Dear student,

You have been selected to participate in a new system at UNED, which wants to economically compensate the effort of students enrolled in the course of “Microeconomía (ADE)”. You have been randomly selected, together with other 138 students, from all the students registered in the course. We remind you that the final exam will take place on Thursday 28th of January of 2016 at 18:30 (first week) and on Thursday 11th of February of 2016 at 11:30 (second week).

You can win **a prize of 1.000€** if your **final grade in the course “Microeconomía I” is above 8.**

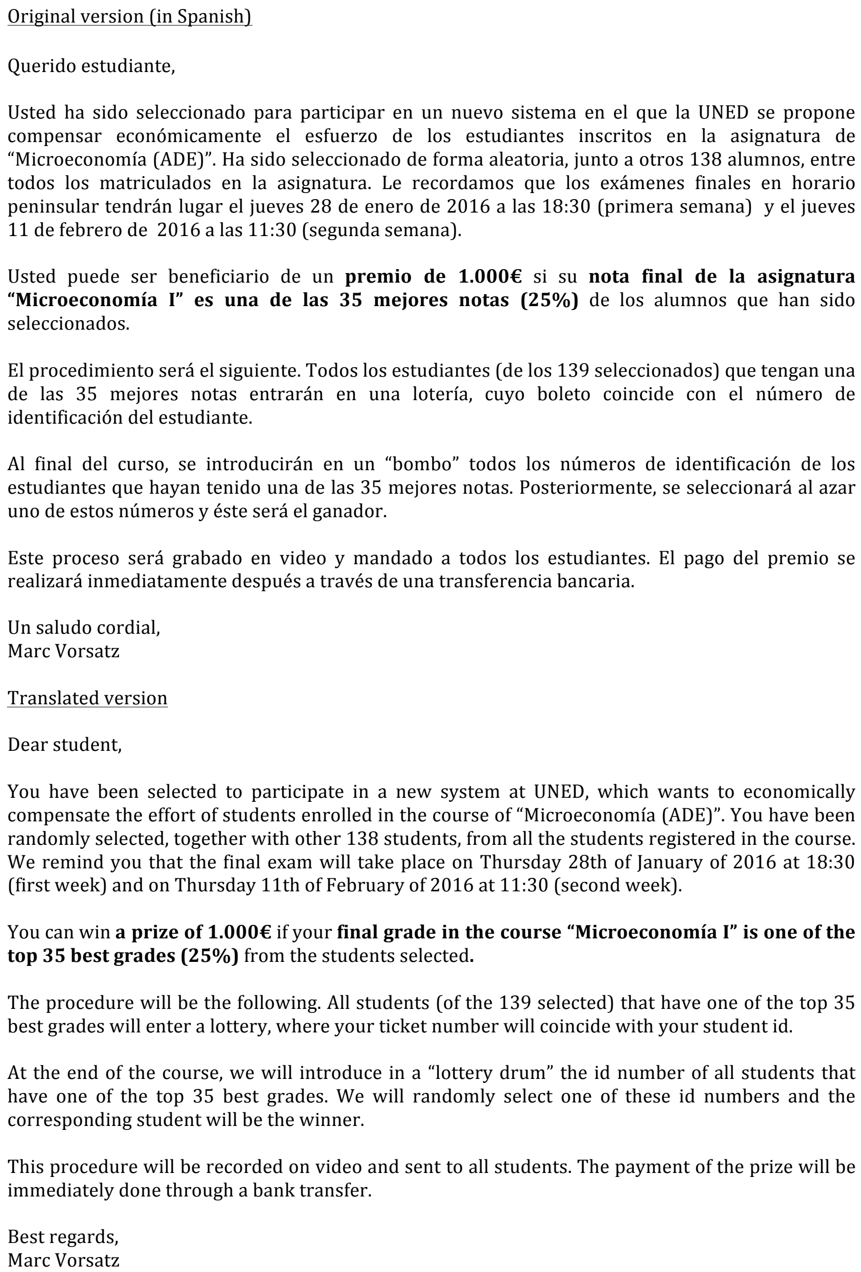
The procedure will be the following. All students (of the 139 selected) that have a final grade above 8 will enter a lottery, where your ticket number will coincide with your student id.

At the end of the course, we will introduce in a “lottery drum” the id number of all students whose grade is above 8. We will randomly select one of these id numbers and the corresponding student will be the winner.

This procedure will be recorded on video and sent to all students. The payment of the prize will be immediately done through a bank transfer.

Best regards,
Marc Vorsatz

Figure 1.A4: EMAIL TREATMENT 2: TOP PERCENTILE



Original version (in Spanish)

Querido estudiante,

Usted ha sido seleccionado para participar en un nuevo sistema en el que la UNED se propone compensar económicamente el esfuerzo de los estudiantes inscritos en la asignatura de “Microeconomía (ADE)”. Ha sido seleccionado de forma aleatoria, junto a otros 138 alumnos, entre todos los matriculados en la asignatura. Le recordamos que los exámenes finales en horario peninsular tendrán lugar el jueves 28 de enero de 2016 a las 18:30 (primera semana) y el jueves 11 de febrero de 2016 a las 11:30 (segunda semana).

Usted puede ser beneficiario de un **premio de 1.000€** si su **nota final de la asignatura “Microeconomía I” es una de las 35 mejores notas (25%)** de los alumnos que han sido seleccionados.

El procedimiento será el siguiente. Todos los estudiantes (de los 139 seleccionados) que tengan una de las 35 mejores notas entrarán en una lotería, cuyo boleto coincide con el número de identificación del estudiante.

Al final del curso, se introducirán en un “bombo” todos los números de identificación de los estudiantes que hayan tenido una de las 35 mejores notas. Posteriormente, se seleccionará al azar uno de estos números y éste será el ganador.

Este proceso será grabado en video y mandado a todos los estudiantes. El pago del premio se realizará inmediatamente después a través de una transferencia bancaria.

Un saludo cordial,
Marc Vorsatz

Translated version

Dear student,

You have been selected to participate in a new system at UNED, which wants to economically compensate the effort of students enrolled in the course of “Microeconomía (ADE)”. You have been randomly selected, together with other 138 students, from all the students registered in the course. We remind you that the final exam will take place on Thursday 28th of January of 2016 at 18:30 (first week) and on Thursday 11th of February of 2016 at 11:30 (second week).

You can win **a prize of 1.000€** if your **final grade in the course “Microeconomía I” is one of the top 35 best grades (25%)** from the students selected**.**

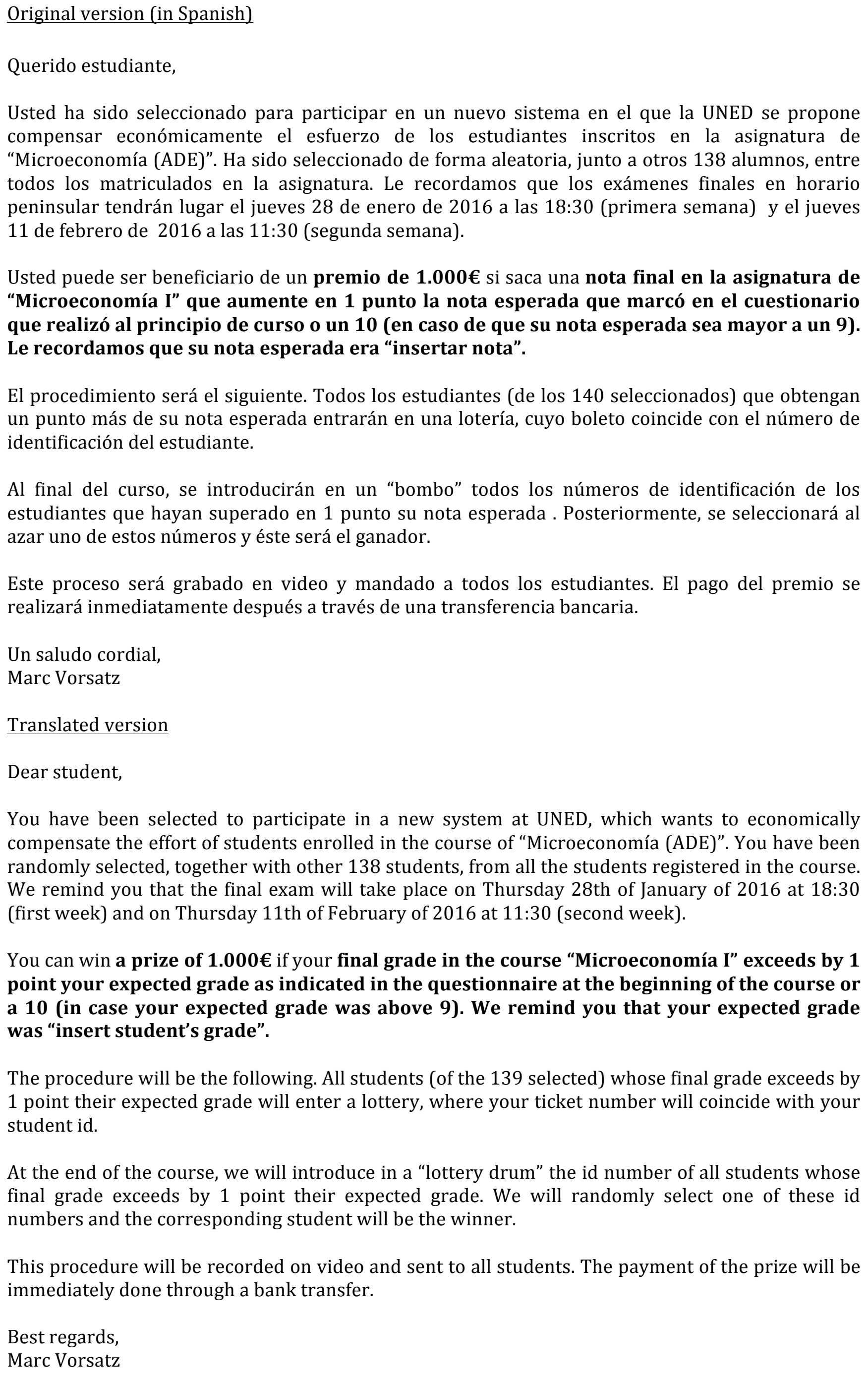
The procedure will be the following. All students (of the 139 selected) that have one of the top 35 best grades will enter a lottery, where your ticket number will coincide with your student id.

At the end of the course, we will introduce in a “lottery drum” the id number of all students that have one of the top 35 best grades. We will randomly select one of these id numbers and the corresponding student will be the winner.

This procedure will be recorded on video and sent to all students. The payment of the prize will be immediately done through a bank transfer.

Best regards,
Marc Vorsatz

Figure 1.A5: EMAIL TREATMENT 3: IMPROVEMENT



Original version (in Spanish)

Querido estudiante,

Usted ha sido seleccionado para participar en un nuevo sistema en el que la UNED se propone compensar económicamente el esfuerzo de los estudiantes inscritos en la asignatura de "Microeconomía (ADE)". Ha sido seleccionado de forma aleatoria, junto a otros 138 alumnos, entre todos los matriculados en la asignatura. Le recordamos que los exámenes finales en horario peninsular tendrán lugar el jueves 28 de enero de 2016 a las 18:30 (primera semana) y el jueves 11 de febrero de 2016 a las 11:30 (segunda semana).

Usted puede ser beneficiario de un **premio de 1.000€** si saca una **nota final en la asignatura de "Microeconomía I" que aumente en 1 punto la nota esperada que marcó en el cuestionario que realizó al principio de curso o un 10 (en caso de que su nota esperada sea mayor a un 9). Le recordamos que su nota esperada era "insertar nota".**

El procedimiento será el siguiente. Todos los estudiantes (de los 140 seleccionados) que obtengan un punto más de su nota esperada entrarán en una lotería, cuyo boleto coincide con el número de identificación del estudiante.

Al final del curso, se introducirán en un "bombo" todos los números de identificación de los estudiantes que hayan superado en 1 punto su nota esperada . Posteriormente, se seleccionará al azar uno de estos números y éste será el ganador.

Este proceso será grabado en video y mandado a todos los estudiantes. El pago del premio se realizará inmediatamente después a través de una transferencia bancaria.

Un saludo cordial,
Marc Vorsatz

Translated version

Dear student,

You have been selected to participate in a new system at UNED, which wants to economically compensate the effort of students enrolled in the course of "Microeconomía (ADE)". You have been randomly selected, together with other 138 students, from all the students registered in the course. We remind you that the final exam will take place on Thursday 28th of January of 2016 at 18:30 (first week) and on Thursday 11th of February of 2016 at 11:30 (second week).

You can win **a prize of 1.000€** if your **final grade in the course "Microeconomía I" exceeds by 1 point your expected grade as indicated in the questionnaire at the beginning of the course or a 10 (in case your expected grade was above 9). We remind you that your expected grade was "insert student's grade".**

The procedure will be the following. All students (of the 139 selected) whose final grade exceeds by 1 point their expected grade will enter a lottery, where your ticket number will coincide with your student id.

At the end of the course, we will introduce in a "lottery drum" the id number of all students whose final grade exceeds by 1 point their expected grade. We will randomly select one of these id numbers and the corresponding student will be the winner.

This procedure will be recorded on video and sent to all students. The payment of the prize will be immediately done through a bank transfer.

Best regards,
Marc Vorsatz

Table 1.A3: Average Effects with No Controls

	Take exam	Midterm	Exam Period	Exam Grade	# questions correct	# questions answered	# correct by # answered	# courses non incentivized	# interventions forum
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Threshold	-0.02	-0.06	-0.04	-0.15	-0.20	0.17	-0.02	-0.17	-0.25
	(0.04)	(0.04)	(0.05)	(0.24)	(0.41)	(0.30)	(0.02)	(0.12)	(0.19)
Top percentile	-0.03	-0.00	-0.01	-0.07	-0.07	0.06	0.00	-0.03	-0.35*
	(0.04)	(0.04)	(0.05)	(0.23)	(0.40)	(0.29)	(0.02)	(0.12)	(0.18)
Improvement	-0.07**	-0.01	-0.02	-0.00	0.16	0.59**	-0.02	-0.17	-0.17
	(0.04)	(0.04)	(0.05)	(0.25)	(0.42)	(0.29)	(0.02)	(0.12)	(0.20)
Observations	1,079	1,079	801	801	801	801	801	1,079	1,079
R^2	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Controls	NO	NO	NO	NO	NO	NO	NO	NO	NO
Mean dep. variable	0.775	0.487	0.367	5.839	12.61	15.99	0.782	1.401	0.674
Std. dev. dep. variable	0.418	0.501	0.483	2.388	4.094	2.952	0.188	1.435	2.626

Notes: The dependent variables are (1) a dummy that takes value 1 if the student has taken the final exam, and 0 otherwise, (2) a dummy that takes value 1 if the student has done the voluntary midterm, and 0 otherwise, (3) a dummy that takes value 1 if the student attended the second period of exams, and 0 otherwise (4) students' exam grade, (5) the number of questions the student answered correctly in the exam, (6) the total number of answered questions in the exam, (7) the number of correct questions divided by the number of answered questions in the exam, (8) the number of courses (non-incentivized) the student has taken during the same term, and (9) total number of interventions in the forum (after the announcement of the treatments) . "Threshold" is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. "Top percentile" is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. "Improvement" is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A4: Heterogeneous Effects by the Willingess to Work of the Students

	Take exam (1)	Midterm (2)	Exam Period (3)	Exam Grade (4)	# questions correct (5)	# questions answered (6)	# correct by # answered (7)	# courses non incentivized (8)	# interventions forum (9)
Threshold	0.00	-0.03	-0.05	-0.13	-0.14	0.11	-0.01	-0.26	-0.42
	(0.04)	(0.05)	(0.05)	(0.24)	(0.41)	(0.30)	(0.02)	(0.16)	(0.28)
Top percentile	-0.02	-0.02	0.05	-0.13	-0.16	-0.08	-0.00	-0.11	-0.49*
	(0.04)	(0.06)	(0.06)	(0.23)	(0.39)	(0.31)	(0.02)	(0.16)	(0.28)
Improvement	-0.03	-0.02	-0.01	-0.07	0.08	0.54*	-0.02	-0.23	-0.26
	(0.04)	(0.06)	(0.05)	(0.25)	(0.43)	(0.30)	(0.02)	(0.16)	(0.31)
Threshold *Willing work	-0.10	0.00	0.06	0.44	0.52	0.41	0.01	0.44	0.71
	(0.09)	(0.12)	(0.12)	(0.55)	(0.94)	(0.73)	(0.04)	(0.30)	(0.64)
Top percentile *Willing work	-0.05	0.12	-0.12	0.29	0.42	0.51	0.01	0.27	0.28
	(0.08)	(0.11)	(0.11)	(0.48)	(0.81)	(0.67)	(0.04)	(0.29)	(0.54)
Improvement *Willing work	-0.12	0.14	-0.05	0.24	0.25	0.17	0.01	0.42	0.21
	(0.09)	(0.12)	(0.12)	(0.54)	(0.92)	(0.69)	(0.04)	(0.32)	(0.53)
Willing work	0.08	-0.10	-0.08	-0.46	-0.49	-0.39	-0.01	-0.59**	-0.84
	(0.07)	(0.10)	(0.10)	(0.46)	(0.75)	(0.56)	(0.04)	(0.28)	(0.63)
Observations	1,079	1,079	801	801	801	801	801	1,079	1,079
R^2	0.09	0.09	0.03	0.27	0.28	0.15	0.20	0.08	0.04
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES
Mean dep. variable	0.775	0.487	0.367	5.839	12.61	15.99	0.782	1.401	0.674
Std. dev. dep. variable	0.418	0.501	0.483	2.388	4.094	2.952	0.188	1.435	2.626

Notes: The dependent variables are (1) a dummy that takes value 1 if the student has taken the final exam, and 0 otherwise, (2) a dummy that takes value 1 if the student has done the voluntary midterm, and 0 otherwise, (3) a dummy that takes value 1 if the student attended the second period of exams, and 0 otherwise (4) students' exam grade, (5) the number of questions the student answered correctly in the exam, (6) the total number of answered questions in the exam, (7) the number of correct questions divided by the number of answered questions in the exam, (8) the number of courses (non-incentivized) the student has taken during the same term, and (9) total number of interventions in the forum (after the announcement of the treatments) . "Willing to work" is a dummy variable that takes value 1 if the student is not working but is willing to work and zero otherwise. "Threshold" is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. "Top percentile" is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. "Improvement" is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, the number of courses students have done before in UNED, and the year the experiment took place. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A5: Non Participant Students

Dependent variables	Control group		Non participants		Diff.
	Mean	St. Dev.	Mean	St. Dev	
# previous courses	5.25	3.67	5.13	3.51	0.12 (0.49)
Mean grade previous courses	6.70	1.00	6.42	0.87	0.28*** (3.75)
Midterm	0.48	0.50	0.26	0.44	0.23*** (6.85)
Take exam	0.78	0.42	0.89	0.31	-0.12*** (-4.37)
Exam grade	5.83	2.40	5.29	2.43	0.54** (2.94)
Observations	263		1250		1513

Notes: The dependent variables are (1) a dummy that takes value 1 if the student has taken the final exam, and 0 otherwise, and (2) the final exam grade. We report the mean, the standard deviation and the difference between the dependent variables for students that participate in the experiment but are assigned to the control group, and for students that did not fill out the pre-experimental questionnarie and did not participate in the experiment.* significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A6: Robustness Check: Heterogeneous Effects by Intrinsic Motivation

	Exam grade				
	< median (1)	> median (2)	1^{st} tercile (3)	2^{st} tercile (4)	3^{st} tercile (5)
Threshold	-0.44	0.38	-0.29	-0.30	0.80
	(0.30)	(0.32)	(0.32)	(0.38)	(0.52)
Top percentile	-0.71**	0.51*	-0.59**	0.05	0.70
	(0.27)	(0.30)	(0.29)	(0.38)	(0.44)
Improvement	-0.32	0.20	-0.19	-0.22	0.37
	(0.31)	(0.32)	(0.35)	(0.36)	(0.50)
Observations	384	417	322	284	195
R^2	0.26	0.27	0.27	0.25	0.32
Controls	YES	YES	YES	YES	YES
Mean dep. variable	5.839	5.839	5.839	5.839	5.839
Std. dev. dep. variable	2.388	2.388	2.388	2.388	2.388

	Exam grade	
	Year 1 (1)	Year 2 (2)
Threshold	-3.04	-1.62
	(2.05)	(2.12)
Top percentile	-3.55*	-4.47**
	(1.96)	(1.96)
Improvement	-2.73	-0.85
	(2.25)	(2.30)
Threshold* Grade satisfaction	0.36	0.19
	(0.26)	(0.28)
Top percentile* Grade satisfaction	0.44*	0.59**
	(0.26)	(0.26)
Improvement* Grade satisfaction	0.33	0.12
	(0.29)	(0.30)
Grade satisfaction	0.24	0.19
	(0.21)	(0.20)
Observations	367	434
R^2	0.09	0.06
Controls	YES	YES
Mean dep. variable	6.267	5.490
Std. dev. dep. variable	2.352	2.370

Notes: The dependent variable is the final exam grade. In the first graph the sample is divided by (1-2) the median, and (3-5) terciles of the number of courses passed before. In the second graph the sample is divided by (1) the first year of the experiment (academic year 2014-5), and (2) the second year of the experiment (academic year 2015-6). "Grade satisfaction" is the grade that the student will feel satisfied obtaining in the course. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, the number of courses students have done before in UNED, and the year the experiment took place. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A7: Robustness Check: Heterogeneous Effects by Experience

	Take exam				
	$<$ median	$>$ median	1^{st} tercile	2^{st} tercile	3^{st} tercile
	(1)	(2)	(3)	(4)	(5)
Threshold	-0.10*	0.04	-0.12	-0.08	0.13**
	(0.06)	(0.04)	(0.07)	(0.06)	(0.05)
Top percentile	-0.01	-0.07	-0.07	-0.09	0.04
	(0.06)	(0.05)	(0.07)	(0.05)	(0.06)
Improvement	-0.14**	0.01	-0.15**	-0.10*	0.05
	(0.06)	(0.05)	(0.07)	(0.06)	(0.06)
Observations	539	540	378	356	345
R^2	0.05	0.08	0.06	0.09	0.09
Controls	YES	YES	YES	YES	YES
Mean dep. variable	0.775	0.775	0.775	0.775	0.775
Std. dev. dep. variable	0.418	0.418	0.418	0.418	0.418

	Take exam	
	Year 1	Year 2
	(1)	(2)
Threshold	-0.15	-0.19**
	(0.11)	(0.09)
Top percentile	-0.08	-0.16
	(0.10)	(0.10)
Improvement	-0.27**	-0.14
	(0.10)	(0.09)
Threshold* # courses before	0.03*	0.03**
	(0.02)	(0.01)
Top percentile* # courses before	0.01	0.02
	(0.02)	(0.01)
Improvement* # courses before	0.05***	0.01
	(0.02)	(0.01)
# courses before	0.01	0.02*
	(0.01)	(0.01)
Observations	503	576
R^2	0.07	0.09
Controls	YES	YES
Mean dep. variable	6.267	0.803
Std. dev. dep. variable	2.352	0.399

Notes: The dependent variable is a dummy that takes value 1 if the student has taken the final exam, and 0 otherwise. In the first table the sample is divided by (1-2) the median, or (3-5) terciles of the number of courses passed before. In the second table the sample is divided by (1) the first year of the experiment (academic year 2014-5), and (2) the second year of the experiment (academic year 2015-6). "# of courses before" accounts for the number of courses that each student has passed before in UNED. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, the number of courses students have done before in UNED, and the year the experiment took place. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A8: Heterogeneous Effects by Previous University Studies

	Take exam (1)	Exam grade (2)
Threshold	-0.04	-0.54*
	(0.05)	(0.28)
Top percentile	-0.05	-0.30
	(0.05)	(0.27)
Improvement	-0.08	-0.39
	(0.05)	(0.30)
Threshold* Previous university studies	0.04	1.21***
	(0.07)	(0.44)
Top percentile* Previous university studies	0.07	0.52
	(0.07)	(0.41)
Improvement* Previous university studies	0.05	0.85*
	(0.07)	(0.46)
Previous university studies	0.01	0.24
	(0.05)	(0.31)
Observations	1,079	801
R^2	0.09	0.27
Controls	YES	YES
Mean dep. variable	0.775	5.839
Std. dev. dep. variable	0.418	2.388

Notes: The dependent variables are (1) a dummy that takes value 1 if the student has taken the final exam, and 0 otherwise, and (2) the exam grade. "Previous university studies" is a dummy that takes value 1 if the student has some previous university studies, and 0 otherwise. "Threshold" is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. "Top percentile" is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. "Improvement" is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, the number of courses students have done before in UNED, and the year the experiment took place. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A9: Heterogeneous Effects by Previous Mean Grade

	Take exam	Exam grade	# questions answered	Perc. questions correct
	(1)	(2)	(3)	(4)
Threshold	0.15	0.32	2.38	-0.02
	(0.27)	(1.51)	(1.99)	(0.13)
Top percentile	-0.27	1.37	5.91***	-0.01
	(0.26)	(1.40)	(1.70)	(0.13)
Improvement	-0.30	-1.12	2.57	-0.15
	(0.27)	(1.41)	(1.86)	(0.13)
Threshold * Mean previous grades	-0.02	-0.01	-0.31	0.01
	(0.04)	(0.22)	(0.28)	(0.02)
Top percentile * Mean previous grades	0.04	-0.17	-0.84***	0.01
	(0.04)	(0.20)	(0.24)	(0.02)
Improvement * Mean previous grades	0.04	0.17	-0.32	0.02
	(0.04)	(0.20)	(0.26)	(0.02)
Mean previous grades	0.02	1.00***	1.12***	0.05***
	(0.03)	(0.15)	(0.18)	(0.01)
Observations	822	660	660	660
R^2	0.07	0.39	0.23	0.27
Controls	YES	YES	YES	YES
Mean dep. variable	0.775	5.839	15.99	0.782
Std. dev. dep. variable	0.418	2.388	2.952	0.188

Notes: The dependent variables are (1) a dummy that takes value 1 if the student has taken the final exam, and 0 otherwise, (2) the exam grade, (3) total number of questions answered in the exam, and (4) the total number of correct divided by the total number of questions answered. "Mean previous grades" is the mean grade of the courses students have passed before in UNED. "Threshold" is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. "Top percentile" is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. "Improvement" is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, the number of courses students have done before in UNED, and the year the experiment took place. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis.* significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A10: Robustness Check: Heterogeneous Effects by Predicted Grade

	# question answered								
	< median	> median	1^{st} tercile	2^{st} tercile	3^{st} tercile	1^{st} quartile	2^{st} quartile	3^{st} quartile	4^{st} quartile
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Threshold	0.37	-0.05	0.30	0.54	-0.46	0.12	0.57	1.02*	-0.38
	(0.40)	(0.37)	(0.49)	(0.52)	(0.44)	(0.59)	(0.53)	(0.61)	(0.51)
Top percentile	0.69*	-0.61*	0.97**	-0.34	-0.89**	0.71	0.09	-0.09	-0.90*
	(0.39)	(0.35)	(0.46)	(0.47)	(0.53)	(0.60)	(0.53)	(0.59)	(0.46)
Improvement	1.41***	-0.22	1.77***	0.67	-0.77*	1.60***	0.79	0.26	-0.44
	(0.39)	(0.37)	(0.49)	(0.50)	(0.51)	(0.58)	(0.57)	(0.53)	(0.50)
Observations	401	400	264	264	273	200	231	180	190
R^2	0.12	0.21	0.15	0.16	0.25	0.18	0.14	0.32	0.29
Controls	NO	NO	NO	NO	NO	NO	NO	NO	NO
Mean dep. variable	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99
Std. dev. dep. variable	2.952	2.952	2.952	2.952	2.952	2.952	2.952	2.952	2.952

Notes: The dependent variable is the total number of questions answered by the student. The sample is divided by (1-2) the median, (3-5) terciles, and (6-9) quartiles of the predicted grade. "Threshold" is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. "Top percentile" is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. "Improvement" is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis.* significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A11: Robustness Check: Heterogeneous Effects by Predicted Grade

	# answered questions	
	Year 1	Year 2
	(1)	(2)
Threshold	5.42**	3.29
	(2.34)	(2.16)
Top percentile	5.61**	5.32**
	(2.24)	(2.06)
Improvement	6.48***	5.43***
	(2.38)	(1.99)
Threshold* Predicted grade	-0.86**	-0.53
	(0.37)	(0.38)
Top percentile* Predicted grade	-0.90**	-0.94**
	(0.35)	(0.36)
Improvement* Predicted grade	-1.02***	-0.81**
	(0.38)	(0.36)
Predicted grade	1.33***	1.10***
	(0.23)	(0.26)
Observations	367	434
R^2	0.07	0.06
Controls	YES	YES
Mean dep. variable	16.44	15.61
Std. dev. dep. variable	2.752	3.067

Notes: The dependent variable is the total number of questions answered by the student during (1) the first year of the experiment (academic year 2014-5), and (2) the second year of the experiment (academic year 2015-6). “Threshold” is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. “Predicted grade” is the prediction of the grade the students under the incentive schemes would have obtained in the absence of the incentive. This predicted grade is calculated following the methodology used by Angrist and Lavy (2009). “Top percentile” is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. “Improvement” is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, and the number of courses students have done before in UNED. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis.* significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.A12: ROBUSTNESS CHECK: ALL HETEROGENEOUS EFFECTS

	Take exam	Exam grade	# questions answered
	(1)	(2)	(3)
Threshold	0.21	-0.41	4.06*
	(0.26)	(1.60)	(2.12)
Top percentile	-0.05	-2.72*	5.44***
	(0.26)	(1.46)	(1.99)
Improvement	-0.42	0.35	6.85***
	(0.27)	(1.76)	(2.15)
Threshold * Grade satisfaction	0.01	0.27	-0.27
	(0.03)	(0.20)	(0.26)
Top percentile * Grade satisfaction	-0.00	0.51***	-0.24
	(0.03)	(0.18)	(0.24)
Improvement * Grade satisfaction	0.01	0.15	-0.21
	(0.03)	(0.22)	(0.27)
Threshold * # courses before	0.03***	0.01	-0.01
	(0.01)	(0.07)	(0.08)
Top percentile * # courses before	0.01	0.03	0.12
	(0.01)	(0.06)	(0.08)
Improvement * # courses before	0.02*	-0.09	-0.10
	(0.01)	(0.07)	(0.08)
Threshold *Predicted grade	-0.08**	-0.31	-0.29
	(0.04)	(0.23)	(0.29)
Top percentile *Predicted grade	-0.00	-0.25	-0.73***
	(0.04)	(0.20)	(0.27)
Improvement *Predicted grade	0.03	-0.18	-0.70**
	(0.04)	(0.23)	(0.28)
Observations	1,079	801	801
R^2	0.11	0.22	0.16
Controls	YES	YES	YES
Mean dep. variable	0.775	5.839	15.99
Std. dev. dep. variable	0.418	2.388	2.952

Notes: The dependent variables are: (1) a dummy that takes value 1 if the student has taken the final exam and 0 otherwise, (2) the final exam grade, and (3) the total number of questions answered by the student. “Threshold” is a dummy variable that is equal to one if the student is assigned to the treatment where students have to reach an 8 in order to win the incentive, and 0 otherwise. “Top percentile” is a dummy variable that is equal to one if the student is assigned to the treatment where students need to have one of the top 25% grades in their group to receive the incentive, and 0 otherwise. “Improvement” is a dummy variable that is equal to one if the student is assigned to treatment where students have to increase by 1 point their ex-ante expected grade in order to receive the incentive, and 0 otherwise. All regressions control for the age, gender of the student, the age at which they started in UNED, their working status, their civil status, the education of their parents, the way they finance their studies, the grade students expect to obtain in the course, the grade students will feel satisfied obtaining in the course, the number of courses students have done before in UNED, and the year the experiment took place. The mean and standard deviation of each dependent variable in the control group is reported in the bottom. Robust standard errors in parenthesis.* significant at 10%; ** significant at 5%; *** significant at 1%.

Chapter 2

Teacher Performance Pay and Student Learning: Evidence from a Nationwide Program in Peru

Joint with María Lombardi

2.1 Introduction

Teacher quality is one of the key factors determining student achievement. Individuals exposed to better teachers not only perform better in school (Hanushek and Rivkin, 2010; Rockoff, 2004; Araujo et al., 2016), but are also more likely to attend college and earn higher salaries (Chetty et al., 2014). However, the payment schemes in most educational systems do not provide adequate incentives for excellence in teaching. With relatively flat salary progression, promotion policies rigidly linked to seniority, and lifetime job tenure, these types of compensation policies might discourage high skilled individuals from taking on the teaching profession, and create weak incentives for existing teachers to exert high levels of effort (Bruns et al., 2011). In an attempt to increase teacher motivation, accountability, effort, and ultimately student learning, academics and policymakers have proposed tying teachers' pay to their students' performance. Pay-for-performance programs in education have been implemented in high income countries like United States, England, the Netherlands and Israel,[1] as well as in developing countries such as India, Pakistan, Kenya, China, Chile, Brazil, Mexico, and more recently in Peru. However, the evidence on the effectiveness of teacher incentives is scant and inconclusive; only part of these programs have been rigorously evaluated, and those that have been studied often differ in their conclusions.

This paper studies the impact of *Bono Escuela* (BE) on student achievement. BE is a nationwide teacher pay-for-performance program implemented in 2015 in public secondary schools in Peru. The program takes the form of a rank-order tournament in which all Peruvian public secondary schools compete within a group of comparable schools on the basis of their annual performance. Every teacher and the principal in schools ranked in the top 20% within their BE group obtain a fixed payment amounting to over a month's salary. The incentives provided by BE are collective (at the school-level), as all teachers are rewarded if their school wins, although the main performance measure used to rank schools is their average score in the 2015 nationwide math and language standardized tests, taken only by 8^{th} graders. This feature of the program, which we exploit in our identification, implies that a school's probability of obtaining the bonus hinges on the achievement of 8^{th} grade students in 2015. Our estimation relies on a novel administrative database collected by the Peruvian Ministry of Education, which covers the universe of students in 2013-2015 and contains annual information on the grades that students receive from their teachers in each subject (their "internal grades").[2] Importantly, teachers' grading tactics should not be influenced by the incentive, since internal grades have no direct impact on a school's BE score. We provide ample evidence that internal grades are correlated with standardized measures of learning, and show that the same teacher typically grades students from parallel classes differently, suggesting that grading on a curve is not the norm in Peruvian secondary schools. The availability of achievement measures for students in all grades allows us to compare changes in the internal grades of

[1] An exhaustive list of OECD countries with teacher pay-for-performance programs is provided in OECD Education at a Glance 2011, available at https://www.oecd.org/education/skills-beyond-school/48631582.pdf

[2] We borrow this terminology from Calsamiglia and Loviglio (2016).

8^{th} graders to those of 9^{th} grade students attending the same school, before and after the incentive was introduced, providing difference-in-differences estimates of the effect of BE on student achievement.

While providing teachers with extrinsic monetary incentives could encourage them to exert more effort, positively impacting student performance, these incentives might yield no improvement if the incentives are not large enough, not understood, or if teachers do not know how to increase student achievement, for example. Teacher incentives might also be ineffective, or even detrimental to student learning, if they lead teachers to engage in undesirable practices such as targeting topics likely to be tested, coaching students on test-taking strategies, or cheating.[3] Student learning could also decrease if the program crowds out teachers' intrinsic motivation.[4] Since a school's probability of obtaining the BE bonus depends on the effort of many of its teachers, the program has the potential of inducing teacher free-riding (Holmstrom, 1982), thus lowering its impact on student learning. However, rewarding the entire school might have the benefit of promoting higher cooperation and monitoring among teachers (Kandel and Lazear, 1992; Kandori, 1992).

We find that the program had no impact on students' math and language internal grades. Our coefficients are precisely estimated, allowing us to reject effects larger than 0.017 standard deviations (SD) in both math and language, well below the treatment effects found in the existing literature (around 0.15-0.25 SD). Furthermore, when separately examining the impact of the program in each of the 395 groups in which schools compete, we find zero average effects in the majority of cases, providing additional evidence of the null effect of BE

[3] This type of behavior is consistent with models of multi-tasking (Holmstrom and Milgrom, 1991; Baker, 1992, 2002), and has been reported in several studies on teacher incentives such as Jacob and Levitt, 2003, Figlio and Winicki, 2005, Figlio, 2006, Glewwe et al., 2010, and Behrman et al., 2015, to name a few. Although these actions could still improve the performance of students whose teachers were devoting little time to effective teaching, they might not affect student learning, or even harm it if they crowd-out effective instruction time (Koretz, 2002; Neal, 2011).

[4] Previous research shows that monetary incentives produce not only a price effect, making the incentivized behavior more attractive, but also a psychological effect that can crowd out the former. There is evidence of this behavior, albeit in a different context, in the studies of Gneezy and Rustichini (2000a), Gneezy and Rustichini (2000c) and Gneezy et al. (2011).

on student achievement. Using data on the overlap of teachers in 8^{th} and 9^{th} grade, we assess whether BE generated improvements in student achievement in our comparison group, and discard the existence of any spillovers which could bias our estimates. Our null average treatment effects could be masking the fact that teachers from some schools might be more incentivized than other, due to the tournament nature of BE. In particular, the incentive could be impacting schools closer to the margin, and leaving sure-winners and sure-losers unaffected (Contreras and Rau, 2012). We explore whether there are differential effects across these and other dimensions of the incentive, and do not find evidence of heterogeneous effects.

Why was student learning unaffected by the BE program? In order to answer this question, we carried out an online survey on a sample of public secondary school teachers regarding the 2015 BE. We provide suggestive evidence that the null effect is not a result of the size or collective nature of the incentive, or driven by teachers being uninformed about the BE, or only focusing on increasing standardized test scores -the incentivized outcome- without influencing their students' learning in a meaningful way. We put forth a few reasons why the program may have had a null effect. Firstly, certain features of the standardized test linked to the bonus might have hampered teachers' ability to boost student performance in terms of this measure, potentially discouraging teachers from exerting higher effort. Given that students were tested for the first time in 2015, teachers might not have known what pedagogical practices result in higher test scores. The fact that students have no stakes in these evaluations might also have played a role in weakening the mapping between teachers' effort and their chances of winning the bonus. Secondly, the incentive might have been diluted if schools were uncertain about their potential ranking within the group of schools they were competing against. Given that they had no prior experience with the standardized test tied to BE, this is not unlikely. Finally, we argue that teachers might not have had enough time to react to the incentive. The future analysis of the 2016 wave of the BE, for which some of these issues will be alleviated, will also allow us to better pin-down these channels.

Our paper relates to the literature on teacher pay-for-performance, particularly in the context of other developing countries. Although a few studies find positive and significant effects on student learning, the literature reveals mixed results. Using a randomized controlled trial in rural schools in the Indian state of Andhra Pradesh, Muralidharan and Sundararaman (2011) study the effect of providing individual and collective monetary incentives to teachers based on students' test score improvements. The incentive had a significant and sizable effect on students' standardized test scores and a positive impact on other subjects not targeted by the incentive. The experimental study of Glewwe et al. (2010) evaluates a collective teacher incentive program in Kenya, and finds that although the program yielded a positive effect on students' test scores in the exams tied to the incentive, it had no impact on non-incentivized exams covering similar topics.[5] Behrman et al. (2015) implement a randomized controlled trial in a sample of Mexican high schools, providing monetary incentives to teachers and/or students based on the latter's performance in math tests with very low stakes. The authors find that, while providing monetary incentives to teachers had no impact on students' math test scores, there was a significant increase in student performance when students themselves were incentivized. The effects were larger when both students and teachers were given incentives.[6] Barrera-Osorio and Raju (2017) evaluate a performance-pay program in Pakistan giving bonuses to primary school principals and/or teachers according to their school's increase in enrollment rates, and the participation rate and scores of their students in a standardized test. The authors only find an increase in the exam participation rates, and argue that this is consistent with the incentives introduced by the program. The closest paper to

[5] Although direct observation of teachers in Muralidharan and Sundararaman (2011) shows no impact of the teacher incentive on classroom processes, or student and teacher attendance, teachers in treatment schools were more likely to report having assigned extra homework, classwork and practice tests, conducting extra classes, and paying special attention to weaker students. External observers in Glewwe et al. (2010) also found no changes in teacher attendance, homework assignment or pedagogy. However, the principals of treated schools were more likely to report that their teachers offered extra prep classes, suggesting that teachers' efforts might have narrowly targeted the incentivized outcome.

[6] Consistent with the authors' findings, incentivized students reported exerting higher effort in preparing for the exam. Self-reported behavior of teachers is not as compatible with the results, since teachers in all three treatment arms were more likely to report having prepared their students for the test, both inside and outside of class.

ours is that of Contreras and Rau (2012), who examine the impact of a scaled-up program in Chile. Using matching and difference-in-differences techniques, they find that public school students performed significantly better in math and language as compared to students in private schools, which were not eligible for the bonus.

In the context of a high income country, the quasi-experimental studies of Lavy (2002) and Lavy (2009) in Israeli high schools examine a collective and an individual teacher incentive program, respectively, and find positive and significant impacts in different measures of student performance tied to the incentives. In a follow-up paper, Lavy (2015) finds that ten years after the pay-for-performance program examined in Lavy (2009), students from treated schools exhibited significantly higher level of schooling attainment and higher wages. Fryer (2013), and Goodman and Turner (2013) independently analyze a randomized controlled experiment in over 200 New York City public schools where schools meeting their performance target could earn a lump sum payment, which they could distribute at their own discretion. Both studies find no evidence of increased student attainment or changes in students' or teachers' behavior. Finally, Springer et al. (2010) conducted a three-year study in the Metropolitan Nashville School System in which math teachers were economically incentivized for large gains on standardized tests, and find a positive effect only among teachers instructing the same set of students in multiple subjects.

Performance pay schemes have traditionally been examined in the context of organizations.[7] There have been several studies examining the causal effect of linking managerial pay to overall firm performance (Groves et al., 1994, Chevalier and Ellison, 1997, and Oyer, 1998, among others) or to the productivity of bottom-tier workers (Bandiera et al., 2007). Other papers have focused instead on the impact of different payment schemes on worker productivity (e.g., Bandiera et al., 2005 and Bandiera et al., 2013). The results of our study can be informative for this other stream of the literature as well.

[7] See Prendergast (1999) for a general review of the early empirical evidence on the provision of incentives in firms.

One contribution of our paper is that we examine whether teacher pay-for-performance can work in the context of a scaled-up, national intervention. Except for Contreras and Rau (2012), all the other studies in this literature tackle this question using a randomized controlled trial, in which the scale is necessarily smaller, and responsibility for the implementation is usually handed over to an NGO (instead of the government). For instance, Barrera-Osorio and Raju (2017) evaluate a pay-for-performance program implemented in 450 schools,[8] Muralidharan and Sundararaman (2011) in 200 schools, Glewwe et al. (2010) in 50 schools, and Behrman et al. (2015) in 40 schools. In contrast, the BE program was implemented solely by the Peruvian Ministry of Education, and reached more than 8,000 schools across Peru, providing incentives to roughly 81,000 teachers, responsible for instructing 70% of Peruvian students in the 8^{th} grade. While these experimental studies make important contributions towards understanding whether teacher pay-for-performance can increase student achievement, they face external validity issues, as in any randomized controlled trial (Deaton and Cartwright, 2016), making their findings not necessarily generalizable to a large-scale program. This notion is put forward in Banerjee et al. (2016), where a successful educational intervention led by a NGO did not yield the same initial impact when it was scaled-up and implemented within the existing educational system. Budgetary constraints (Kerwin et al., 2015) and opposition from teacher unions (Bruns and Luque, 2015; Mizala and Schneider, 2014) make several aspects of these types of interventions unfeasible in a nationwide program. For example, students in Muralidharan and Sundararaman (2011) were evaluated at baseline, and their teachers received feedback on their performance in each question. Testing students so often and providing such detailed feedback to their teachers might be too costly to implement on a national scale. It is thus crucial for policymakers to better understand the role played by the features of teacher pay-for-performance programs. While we cannot fully tease out which characteristics of the BE contributed to its null impact, we provide some suggestive evidence, hopefully shedding more light on this discussion.

[8] Although in comparison to BE this program was implemented in a reduced sample of schools, it shares the feature that it was designed and managed by the government.

Another novelty of our study is its use of a measure of student achievement that captures the skills of students which are targeted by the program without being directly incentivized. Since the BE bonus is linked to standardized test scores and not to internal grades, teachers' stakes in our outcome variable are not modified by the incentive.[9] An identification strategy relying on standardized test scores (or other incentivized indicators) as an outcome cannot fully disentangle whether improvements in students' performance are the consequence of higher learning or the results of short-term strategies fostering high test scores (Neal, 2011). The importance of this issue is highlighted by the results from Glewwe et al. (2010), who find that while students performed better in the tests used to award the bonus, there was no effect in their performance in an alternative exam not linked to the incentive. With the exception of the latter, all the other papers in this literature assess student achievement using measures of learning which are directly targeted by the incentive.[10] While it would also be interesting to analyze the impact of BE on standardized test scores, it is not possible because 2015 was the first year in which these tests were applied in secondary schools, and there is no appropriate comparison group. Using internal grades as our outcome has the advantage of capturing students' performance without directly influencing teachers' probability of obtaining the bonus. While this measure might have some shortcomings, for instance if teachers assign grades on a relative basis, we report the results from multiple tests alleviating this concern.

The paper is organized as follows. Section 2.2 describes the educational sys-

[9] Students' internal grades in Peru are completely independent of standardized test scores. For one, standardized tests are graded after the end of the school year, and students' individual scores are never reported.

[10] Behrman et al. (2015) and Contreras and Rau (2012) measure achievement using students' scores in the standardized evaluations tied to the bonus. The studies of Fryer (2013) and Goodman and Turner (2013) use several measures of student performance linked to the incentive (scores in state tests, graduation rates, credits earned, etc.), as do Lavy (2002) and Lavy (2009) (average score and pass rates in matriculation exams, and school dropout rates). In an attempt to overcome this issue, Muralidharan and Sundararaman (2011) designed the standardized test to include mechanical and conceptual questions; while performance in the former can be easily affected by a teacher coaching his students for the test, conceptual questions are harder to influence using these types of tactics.

tem in Peru and the Peruvian teacher pay-for-performance program. Section 2.3 discusses our strategy for estimating the effect of teacher incentives on student performance, and Section 2.4 describes the data. Section 2.5 presents our main results, and Section 2.6 provides evidence on the validity of our identification assumption. Section 2.7 discusses the potential reasons for our findings, and Section 2.8 concludes.

2.2 Secondary Schooling in Peru and the BE Program

2.2.1 Secondary Schooling in Peru

Compulsory schooling is 12 years in Peru, and is composed of initial, primary and secondary schooling, lasting one, six, and five years. Students in public secondary schools have seven hours of instruction a day, although the Ministry of Education has been gradually implementing nine-hour school days, currently reaching 18% of all public secondary schools. While a significant portion of the student body attends private secondary schools, public institutions dominate by far. In 2014, for instance, 63% of high schools were publicly run, instructing 76% of all secondary students. Over the last decade there have been significant improvements in secondary school coverage, with enrollment rising from 71% of individuals in secondary school age (12-16) in 2005 to 83% in 2014. Despite these improvements, enrollment is still far from universal. Moreover, a very high portion of students attending high school do not possess the minimum required levels of knowledge. In the 2012 round of OECD's Programme for International Student Assessment (PISA) evaluating 15-year-old students, Peru was the lowest scoring country out of 65 in all three tested subjects. In particular, 75%, 60% and 69% of Peruvian students had low achievement in math, reading and science, respectively.

Public school teachers in Peru can be either civil servants or contract teachers. Salaries for the former are divided into eight pay scales, with a monthly salary of 1451 soles ($\approx$ 439 dollars) in the lowest scale in 2015, and a salary of 3773 soles ($\approx$ 1142 dollars) in the highest.[11] Contract teachers, on the other hand, received a fixed monthly payment of 1244 soles ($\approx$ 370 dollars).[12] There were approximately 120,000 public secondary school teachers in 2014, one third of which were contract teachers. The average secondary school teacher in public schools received a monthly salary of only 1469 soles, roughly 444 dollars.[13] The working week for public secondary school teachers in 2015 consisted of 26 hours, 24 of which were to be spent teaching.[14] However, as reported in a nationally representative teacher survey at the end of 2014, teachers spent an average of 12 hours a week performing other activities outside their official working hours, such as preparing class materials or attending parent-teacher conferences. Furthermore, 15% of secondary school teachers taught in more than one school, and 28% complemented their salary with another type of job.

According to this same survey, 52% of public secondary school teachers had a university degree, 45% obtained their teaching degree in a tertiary institution, and the remaining 3% had another type of degree, or no degree at all. As compared to Peruvian workers with similar qualifications, and teachers in comparable countries, Peruvian teachers are poorly paid. A study on Latin American teachers' salaries in 2010 shows that adjusting for the number of hours worked, Peruvian teachers made 10% less than other Peruvian professional workers with similar education (Bruns and Luque, 2015). In comparison to individuals with similar qualifications, teachers in Peru were paid relatively worse than in Mexico, Honduras, El Salvador, Costa Rica, Uruguay and Chile, but relatively better than than in Panama, Brazil and Nicaragua.[15] While Peruvian teachers are poorly paid, ab-

[11] Throughout this study we use a conversion rate of 3.31 soles per dollar.

[12] Further details on teachers' salaries and pay scales are provided by the Ministry of Education in http://www.minedu.gob.pe/reforma-magisterial/remuneraciones-beneficios.php, last accessed August 16, 2016.

[13] We calculated the average monthly salary of public secondary school teachers in Peru using the Ministry of Education's pay scales and the type of contract and category reported by a nationally representative sample of secondary school teachers in a survey conducted by the Ministry of Education at the end of 2014 (*Encuesta Nacional de Docentes*).

[14] In 2015, the working week was expanded by two (paid) hours, which are meant to be spent performing activities outside the classroom, namely preparing materials for class, assisting students who fall behind, providing orientation to parents, etc. In 2016, an extra two hours were added, reaching a total of 30 working hours a week.

[15] Mizala and Ñopo (2016) examine the patterns of teacher pay in several Latin American coun-

senteeism is quite low. Around the time in which BE was implemented, teacher absenteeism was below 7% in public schools around the country.[16]

2.2.2 The BE Program

In 2013, the Peruvian Ministry of Education launched *Bono Escuela* (BE), a nationwide teacher pay-for-performance program in public schools. The program was first implemented in primary schools, and extended to secondary schools in 2015. Secondary schools, the focus of this paper, were only included in BE starting 2015 because Peru's census standardized tests (*Evaluación Censal de Estudiantes*, henceforth ECE), one of the key indicators used for the BE, were not implemented in secondary schools until 2015.[17] The ECE is an annual low-stakes test designed by the Peruvian Ministry of Education, in which students from different grades in practically all private and public schools are tested on their basic competencies in math and language at the end of the school year.[18] In secondary schools, only 8^{th} graders are tested. The ECE is implemented by the Peruvian National Statistics Institute (INEI), which trains independent enumerators for this task. Since the main goal of the ECE is to track the evolution of student learning throughout the country and help shape educational policies, school average scores are reported to school district governments, schools and parents.

tries in an earlier period (1997-2007), and find that teachers in Nicaragua and Peru were the most underpaid relative to their nationals working as professionals or technicians.

[16] Around April 2015, the Ministry of Education launched *Semaforo Escuela*, a program in which trained enumerators make periodic visits to public schools, and register information on teacher, student, and director absenteeism, among other things. Further details are available at http://www.minedu.gob.pe/semaforo-escuela/. In a 2006 study, teacher absenteeism in Peruvian public schools was found to be higher, around 11% (Chaudhury et al., 2006).

[17] We do not examine the effect of the primary school BE program due to identification issues related to the timing of the program's implementation. The BE was first announced by the president of Peru in July 2014, although the corresponding regulation only came out in October of that year. The 2013 edition was implemented retroactively (i.e., after the 2013 ECE test had been taken), in an attempt to boost the program's credibility. In the case of 2014, it is unclear whether schools knew about the program before taking the ECE in November, since the BE was broadly announced four months before the test, but its regulation only came out one month before.

[18] The ECE was first implemented in 2007 in 2^{nd} grade of primary school, and was extended in the following year to 4^{th} grade in schools with intercultural bilingual education. It was administered in 8^{th} grade for the first time in 2015, and will be extended to 4^{th} graders in all schools in 2016.

Besides being an informational tool for the Ministry of Education, ECE test scores are one of the metrics used to rank schools and select the BE bonus recipients. Schools not eligible for taking the ECE test in 2015 (only 4% of all public secondary schools) compete for a smaller bonus based on other measures. We focus our analysis solely on public schools taking the ECE. As outlined in Table 2.1, a school's score for the BE is composed of several factors. The score gives 40% of weight to the average math and language grade of 8^{th} graders in the ECE standardized tests.[19] In order to prevent teachers from encouraging absenteeism of low achieving students on the day of the ECE evaluation, schools not complying with a minimum rate of student participation are disqualified from taking part in the BE. In particular, ECE participation must be 80%, 90% and 95% in schools with only one, two or more than two 8^{th} grade classes[20] Additionally, 35% of weight is given to the entire school's intra-annual retention rate, that is, the proportion of enrolled students still in school at the end of the year. Although dropout rates are non-negligible, most of the dropping out takes places after the school year ends, making retention rates already extremely high before the program was implemented. The average retention rate in the public secondary schools was 99% in 2014, and only 7% of schools had retention rates below 95%. In practice, schools had very little leeway for improving their retention rates, and could thus not compete on the basis of this indicator.[21] An extra 5% of the school's score depends on whether the principal enrolls his students in the Ministry of Education's administrative system (*Sistema de Información de Apoyo a la Gestión de la Institución Educativa*, henceforth SIAGIE) in a timely manner, something which should not affect the incentives of teachers and thus the performance of their students. The remaining 20% of the score depends on an index of school management, composed of teacher attendance, management

[19] In the primary school BE program, the score is also composed of the change in the average ECE scores from the previous year, to incentivize schools in the lower end of the distribution. Since 2015 was the first year the ECE was implemented in secondary schools, this could not be replicated.

[20] 91% of public secondary schools complied with this requirement, and the average school only had 1.4 students absent on the day of the exam.

[21] In practice, giving such a large weight to this indicator counteracts any perverse incentives schools might have to improve their test scores by encouraging their weakest students to drop out

of school infrastructure, compliance with class hours, as well as measures of pedagogical practices and learning environment. The first three measures are collected by independent evaluators making visits to all public schools, whereas the last two are obtained from questionnaires handed out to 8^{th} grade students during the ECE. All in all, around 80% of a school's score ultimately depends on the performance of 8^{th} grade students in math and language. Consistent with this fact, schools ranked in the top 20% of their BE group according to their average ECE score were 57 percentage points more likely to win the bonus.

Table 2.1: Assignment of Score in BE

Weight	**Indicator**	**Relevant Grades**
40%	Average math and language score in 2015 ECE standardized tests	8th Grade
35%	Intra-annual retention rates	All Grades
5%	Enrollment of students in SIAGIE administrative system	All Grades
12%	Teacher attendance, management of school infrastructure and compliance with class hours	All Grades
8%	Pedagogical practices and learning environment	8th Grade

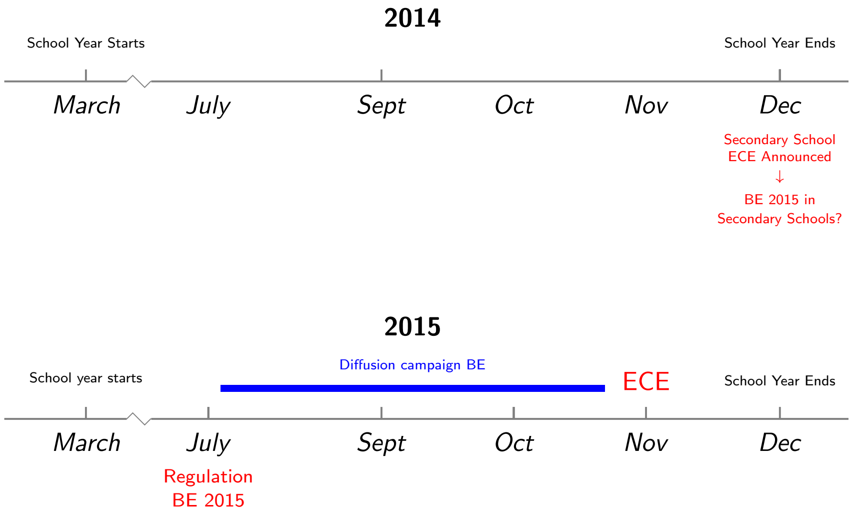
Source: Decree 203-2015.

The timing of the BE is depicted in Figure 2.1. The school year in Peru starts in March, and ends in December. At the end of the 2014 school year, once the implementation of the ECE test in secondary schools was confirmed for the following year, the Minister of Education announced the possibility of extending the BE program to secondary schools as well.[22] The government resolution regulating the 2015 BE came out on the 25^{th} of July, almost four months before

[22] http://larepublica.pe/21-12-2014/jaime-saavedra-el-proceso-para-nombrar-a-8-mil-maestros-se-inicia-en-julio-del-2015 (last accessed August 16, 2016).

the 2015 ECE (carried out in November 17/18), and was accompanied by a diffusion campaign launched by the Ministry of Education informing schools about the BE program. In comparison to other studies, the time frame teachers had to react was relatively short. We elaborate on this issue using the results of our teacher survey in Section 2.7.6.

Figure 2.1: TIMING OF BE IN SECONDARY SCHOOLS



BE is set up as a collective incentive, such that the principal and every teacher in a school are rewarded if the school scores in the top 20%.[23] To ensure that schools competing against each other are comparable, they are separated into groups by school district, instruction time, and by whether they are urban or rural. There are 395 groups in total, with an average of 34.6 schools per group. Teachers in schools in the top 10% in their group get a bonus of 2000 soles (roughly 605 dollars), whereas those in schools in the top 10%-20% get paid 1500 soles (454 dollars). Every teacher in a winning school gets the exact same bonus, whereas

[23] Other papers studying collective teacher incentives are Lavy (2002) in Israel, Glewwe et al. (2010) in Kenya, Muralidharan and Sundararaman (2011) in India, Contreras and Rau (2012) in Chile, and Fryer (2013) and Goodman and Turner (2013) in the US.

the school principal gets a slightly larger payment (500 extra soles). Since the average teacher receives a monthly salary of 1469 soles, the bonus constitutes either 1 or 1.4 monthly salaries on average. Considering that 20% of schools receive the prize, the average value of the bonus is 24% of a monthly salary, a sizable figure as compared to other studies in the literature.[24,25]

2.3 Estimation Strategy

We exploit the fact that a school's score for the BE largely depends on the performance of 8^{th} grade students in the math and language ECE test for estimating the causal effect of the teacher incentive on student learning. This feature of the BE results in schools having a much higher incentive to improve the learning of 8^{th} grade students as compared to students from other grades. With this notion in mind, we perform a difference-in-differences estimation comparing the change in achievement of 8^{th} grade public school students with that of 9^{th} grade students attending the same school.[26] In our preferred specification, we use a repeated cross-section of 8^{th} and 9^{th} grade students in public secondary schools eligible for the BE, and run the following regression:

$$\begin{aligned} Internal\,Grade_{ist} = \beta_0 + \beta_1\, 8^{th}\, Grade_{ist} + \beta_2\, 8^{th}\, Grade_{ist} * Post_t \\ +X_{ist}\,\delta + \gamma_t + \gamma_s + U_{ist} \end{aligned}$$

where $Internal\,Grade_{ist}$ is the grade that student *i* in school *s* and year *t* obtained

[24] The average payment is 350 Soles (0.1x2000 + 0.1x1500), which constitutes 24% of the average teachers' monthly salary.

[25] In the teacher incentive program of Muralidharan and Sundararaman (2011), the average bonus was around 35% of a monthly salary; in the experiment run by Glewwe et al. (2010) in Kenya prizes were in-kind, and the average teacher got a bonus worth 12%-21% of a monthly salary. In the Israeli program studied by Lavy (2002) prizes of 10%-40% of the average teacher's monthly salary were awarded to approximately one third of participating teachers, whereas the prizes roughly represented 50% of a monthly salary and were awarded to two thirds of teachers in the New York experiment studied in Fryer (2013) and Goodman and Turner (2013). The incentive implemented in Chile and studied by Contreras and Rau (2012) awarded an average bonus of 10% of a monthly salary.

[26] We use 9^{th} grade as our comparison group and not 7^{th} grade, for example, because the program might have an impact on teachers in the latter grades, given that their students will be taking the ECE standardized test in 2016.

in a particular subject (i.e., the grade assigned to student *i* by his/her teacher at the end of the school year). We run separate regressions using math and language internal grades in our main specification, and also estimate this equation for the average internal grade in all subjects not evaluated in the ECE, to examine whether the BE impacted students' performance in other courses. $8^{th}\ Grade_{ist}$ is a dummy for whether student *i* from school *s* is an 8^{th} grader in year *t*, $Post_t$ is a dummy taking the value of one in the year 2015 and zero in 2013-2014, X_{ist} is a set of individual controls (gender, if Spanish is the student's native tongue, if the student was retained in the previous year, has a disability, and whether the parents are alive and living in the same household), and γ_t and γ_s are year and school fixed effects. We run regressions for the period 2013-2015, i.e., two years before the BE, and the year in which it took place. Our estimation thus compares students in 8^{th} and 9^{th} grade, within the same school, before and after the BE was introduced. Including school fixed effects allow us to restrict our comparison to students facing the same educational environment, but differing in their exposure to the BE.[27] U_{ist} are all the unobserved determinants of achievement for student *i* in school *s* and year *t*, such as ability, motivation, household income, and home environment, to name a few. We allow for our errors to be correlated within school by clustering our standard errors at the school level. We express grades as a z-score, standardizing them by subject and year, so that our coefficient of interest (β_2) can be interpreted as the standard deviation (SD) change in internal grades associated with the incentive. In the case of non-incentivized courses, we first calculate the z-score for each course, and then take the average. As a robustness check, we also standardize internal grades for each subject by school and year.

Unlike other studies on teacher pay-for-performance, our outcome variable

[27] Given that 8^{th} grade students in private schools take the ECE but these institutions are not eligible for the BE, we could also run a differences-in-differences regression comparing the change of internal grades of 8^{th} grade students from public and private schools, similar to what Contreras and Rau (2012) do for the case of Chile. However, as shown in Appendix Table 2.A1, public school students were already improving relatively faster than their private school counterparts in the year prior to the BE (i.e., the *Public* × *2014* coefficient is statistically significant). Since there are other things that could be changing across the public-private spectrum in 2015 that we cannot control for, we discard this estimation strategy.

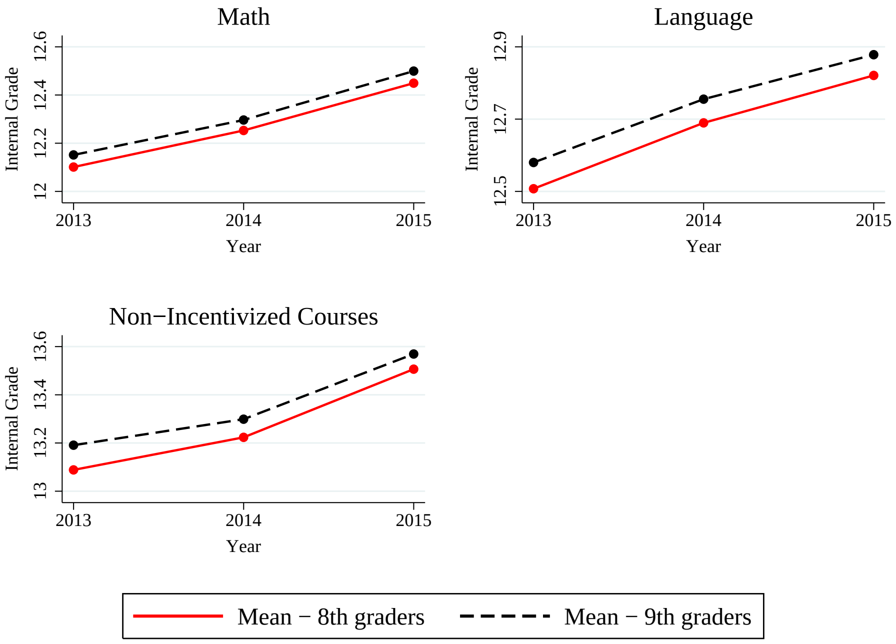
is the grade assigned to students by their teachers at the end of the school year (what we refer to as internal grades), and not their standardized test results. Given that teachers' pay under the BE is tied to performance in the ECE, an identification strategy relying on standardized test scores as an outcome cannot disentangle whether improvements in students' performance are the consequence of increased student learning or the results of short-term strategies fostering high test scores (Neal, 2011). Having internal grades as our outcome has the advantage of capturing students' performance without directly influencing teachers' probability of obtaining the bonus. While it would still be interesting to study the impact of the BE on students' ECE test scores, we cannot do so because the ECE test was applied in secondary schools for the first time in 2015, and there is no group of students serving as an appropriate comparison. From the perspective of students, internal grades play a very important role, directly affecting whether they pass the school year, take summer remedial courses or are retained. Importantly for our identification, teachers' grading tactics should not be influenced by the BE incentive. Since the bonus from the BE is tied to ECE test results, and not internal grades, teachers' stakes in their students' internal grades are not directly modified by the incentive program.[28] Although it is mandatory for schools to report students' internal grades to the Ministry of Education, these grades have absolutely no bearing on whether the school obtains the bonus. As long as internal grades present enough variability, we would expect them to reflect changes in learning. We provide supporting evidence of the fact that within-schools, internal grades are correlated with standardized measures of learning, and show that grading on a curve is uncommon in Peruvian secondary schools in Section 2.6.2.

Our main identifying assumption is that in the absence of the teacher incentive, the performance of 8^{th} and 9^{th} grade students attending the same school would have evolved in an equivalent way between 2014 and 2015. A necessary

[28] Our teacher survey inquires, among other things, about whether teachers changed the difficulty of their classes in 2015 as a result of BE. As shown in Table 2.10, teachers were equally likely to report that they decreased the difficulty of their classes when teaching students from 8^{th} grade, as compared to students from other grades, and only 5 percentage points more likely to report that they increased the difficulty of their classes.

condition for giving a causal interpretation to β_2 is that 8^{th} and 9^{th} grade students follow parallel trends before the implementation of the BE. An inspection of the raw means in Figure 2.2 shows that grades of 8^{th} and 9^{th} grade students appear to be on parallel trends in both math and language before the program was implemented. We provide formal evidence for the parallel trends assumption in Section 2.6.1.

Figure 2.2: TREND IN AVERAGE INTERNAL GRADES FOR 8^{th} AND 9^{th} GRADERS



Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools which were eligible for taking the 2015 ECE standardized test and which are registered in the Ministry of Education's SIAGIE administrative system. The figures plot the average of all 8^{th} and 9^{th} graders internal grades in math, language and non-incentivized courses, respectively. We take the average of non-incentivized courses, which are art, science, social studies, English, civics, human relations, physical education, religion, and education for the workforce.

Identifying a causal effect also requires that the performance of 9^{th} grade students, our comparison group, is unaffected or hardly affected by the teacher incentive program (i.e., that there are no spillovers). Importantly for our identification, schools do not have much room to compete on the basis of indicators other than the 8^{th} graders' standardized test scores, leading to a practically null correspondence between 9^{th} grade students' learning and a school's BE score. As explained in Section 2.2.2, around 80% of a school's score ultimately depends on the performance of 8^{th} grade students. This implies that, if any, a very small portion of the school's score could be improved if 9^{th} grade teachers exerted more effort. It is important to bear in mind, however, that since 83% of 8^{th} grade teachers also instruct 9^{th} grade, an increase in effort while teaching 8^{th} graders could potentially spill over to students in our comparison group and bias our estimates downwards. We show that this is not a concern by exploring the impact of the teacher incentive in schools with a low overlap between 8^{th} and 9^{th} grade teachers in Section 2.6.2. On the other hand, the fact that the probability of obtaining the bonus hinges largely on the performance of 8^{th} grade students might lead the school to redirect its resources towards these grades, negatively impacting the internal grades of students in our comparison group. We discuss this in further detail in Section 2.6.3, and show that this issue is not a concern in our setting.

2.4 Data and Descriptive Statistics

2.4.1 Administrative Data

Our empirical exercise relies on a rich administrative database collected by the Peruvian Ministry of Education in 2013-2015, derived from its SIAGIE system. Coverage is basically universal, reaching 99.7% of public schools. Schools must enroll their students into the SIAGIE system at the start of the school year, and input the final grades of their entire student body once the academic year concludes. Grading is done on a 0-20 scale, and students need to obtain at least 11 to pass a given subject. Besides students' grades, this database also has information on characteristics such as age, gender, native tongue, parents' education, if they live with their parents, etc. Student identifiers permit tracking individual students across years. The SIAGIE also contains information on the grade and classroom that student are assigned to, the teachers who teach each grade and group, and some basic teacher characteristics such as age and gender. In 2015, there were 8,654 public secondary schools in Peru, of which 8,092 were eligible for participating in the ECE. Schools must have at least five 8^{th} grade students in order to be eligible for taking the test. Our SIAGIE database covers 8,059 of these schools.

Table 2.2 presents some characteristics of the 8^{th} and 9^{th} grade students attending public secondary schools in 2013-2015. We observe that the mean final grade in math is 12.27 and 12.32 (out of 20) in 8^{th} and 9^{th} grades, and 84% and 85% of students pass this course. Students perform slightly better in language, where the mean final grade for the 8^{th} and 9^{th} grade students is 12.67 and 12.64, and 89% and 90% of them pass the course. Mean grades in other courses exceed those of math and language by almost one point, and almost all (93% and 94%) students pass these courses. Half of the students are male, almost all of them are natives, and 83-84% of them have Spanish as a native tongue. Only 6% and 4% of 8^{th} and 9^{th} graders were retained in the same grade the year before. Although it is not necessary for our identification that 8^{th} and 9^{th} grade students are balanced in terms of observables, they do appear to be very similar. In addition, Table 2.2 shows some characteristics of the 8,059 public secondary schools in our sample. Less than half (40%) of the public secondary schools are located in rural areas. Each school has, on average, two classes per grade, and there are around 19 students per teacher in the average class. We also observe that each school has, on average, roughly 11 teachers teaching 8^{th} and 9^{th} grade, with 83% of teachers in 8^{th} (9^{th}) grade also teaching 9^{th} (8^{th}) grade. Teachers are almost 42 years old on average, and 60% of them are male.

2.4.2 Survey Data

We complement our main empirical analysis with the results of an online survey we conducted with the assistance of the Ministry of Education. According to our teacher database, there were 123,669 public secondary school teachers in 2015. The Ministry of Education has the email address of 36,283 of them (30%), all of which received a survey email from the Ministry in October 2016, a few weeks before the winners of BE were announced. As in the past

Table 2.2: Summary Statistics for 8^{th} and 9^{th} Graders

	8^{th} Grade		9^{th} Grade	
	Mean	Std. Dev	Mean	Std. Dev
Final Grade (0-20)				
Math	12.27	2.17	12.32	2.17
Language	12.67	2.07	12.74	2.07
Other courses - average	13.27	1.60	13.35	1.60
Passed the Course				
Math	0.84	0.37	0.85	0.36
Language	0.89	0.31	0.90	0.30
Other courses - average	0.93	0.15	0.94	0.14
Other Individual Characteristics				
Male	0.49	0.50	0.50	0.50
Repeated last year	0.06	0.23	0.04	0.20
Foreigner	0.00	0.05	0.00	0.05
Spanish is native tongue	0.84	0.37	0.83	0.38
Has a disability	0.00	0.06	0.00	0.06
Father is alive	0.90	0.30	0.89	0.31
Mother is alive	0.97	0.16	0.97	0.17
Father lives in HH	0.76	0.43	0.77	0.42
Mother lives in HH	0.80	0.40	0.80	0.40
Number of students	1,090,496		1,018,310	
Grade/School Characteristics				
Rural	0.41	0.49	0.40	0.49
Number of classes	2.00	1.92	1.94	1.84
Teacher-pupil ratio	19.61	8.72	18.98	8.75
Number of teachers	10.74	6.55	10.95	6.71
% of teachers instructing the other grade	0.83	0.22	0.83	0.20
Average age of teachers	41.64	5.34	41.66	5.26
% of male teachers	0.60	0.21	0.60	0.20
Number of school-year observations	23,810		23,469	

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. We exclude students for which we have no grades and/or no individual controls (0.4%). Since teacher data is missing for a small subsample of schools, the number of grade-observations for teacher characteristics is 23,462 and 23,127 in 8^{th} and 9^{th} grade. *Final Grade* is the students' internal grades at the end of the school year in math, language and non-incentivized courses. *Passed the Course* is a dummy for whether the student got an 11 or higher in that particular course (the requirement for passing). We take the average of non-incentivized courses, which are art, science, social studies, English, civics, human relations, physical education, religion, and education for the workforce. *Repeated last year* is a dummy for whether the student was retained in the same grade at the end of the previous year. *Rural* is a dummy for whether the school is in a rural area, and *Number of classes* is the number of classes in the student's grade and year. *Number of teachers* is the total number of teachers in that grade and year, and *% of teachers instructing the other grade* is the % of 8^{th} (9^{th}) teachers also teaching 9^{th} (8^{th}) grade in the same school.

editions of BE in primary schools, the bonus winners were announced at the end of the following school year (in November 2016). Teachers were asked what grades and subjects they taught in 2015, their knowledge about the BE and its rules at that time, and their opinion about the size of the bonus. We also inquired about changes in their pedagogical practices while teaching students from different grades, and about administrative changes in the school they were working for in 2015. Finally, we tried to elucidate teachers' perception about their school's ranking and its probability of winning, and asked teachers for their opinion about students' motivation in the standardized test tied to the BE.

The survey was anonymous, and teachers were told that its purpose was to collect information about teachers' perceptions and opinions about the BE program. Since the survey was framed in the context of BE, and sent by the Ministry of Education, respondents might be subject to social desirability bias (i.e., over-reporting of good behavior associated with the objectives of BE). To try and maximize the response rate, and due to restrictions imposed by the Ministry, we did not ask questions about teacher characteristics or identify the school they worked for, and thus we cannot compare survey respondents to non-respondents. We received a response from 3,406 teachers (9.4% response rate), roughly 2.8% of all public secondary school teachers. Given the potential bias in teachers' responses and our selected sample, the results from this survey must be taken with caution.

2.5 Results

The teacher incentive program had no effect on 8^{th} grade students' math and language internal grades, as shown in columns (1) and (4) of Table 2.3. Our coefficient of interest (the interaction of the 8^{th} *Grade* and *Post* dummies) is robust to the inclusion of school fixed effects (columns 2 and 5) and individual controls (columns 3 and 6), with the latter being our preferred specification.[29]

[29] In our baseline regressions we standardize internal grades by subject-year, but the results are quantitatively similar if we standardizing each subject by school and year. Parents' education is missing for 12% of students, so we do not control for this in our baseline regressions. However, attrition is not differential across grades, and our results are robust to controlling for this.

Table 2.3: EFFECT OF TEACHER INCENTIVE ON STUDENTS' MATH AND LANGUAGE INTERNAL GRADES

	Math			Language		
	(1)	(2)	(3)	(4)	(5)	(6)
8^{th} Grade * Post	-0.001	-0.003	-0.005	0.006	0.004	0.001
	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)
8^{th} Grade	-0.022***	-0.016***	-0.008**	-0.034***	-0.026***	-0.016***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Repeated last year			-0.570***			-0.623***
			(0.006)			(0.007)
Male			0.115***			0.334***
			(0.003)			(0.003)
Foreigner			0.046**			0.089***
			(0.019)			(0.019)
Spanish is native tongue			0.171***			0.174***
			(0.007)			(0.007)
Has a disability			-0.262***			-0.256***
			(0.014)			(0.015)
Father is alive			0.066***			0.066***
			(0.004)			(0.003)
Mother is alive			0.040***			0.049***
			(0.006)			(0.006)
Father lives in HH			0.026***			0.020***
			(0.002)			(0.002)
Mother lives in HH			0.014***			0.019***
			(0.003)			(0.002)
Observations	2108806	2108806	2108806	2108793	2108793	2108793
R^2	0.000	0.071	0.092	0.000	0.087	0.135
Year FE	YES	YES	YES	YES	YES	YES
School FE	NO	YES	YES	NO	YES	YES
Individual Controls	NO	NO	YES	NO	NO	YES

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. The dependent variables are students' internal grades in math and language, standardized by course-year. 8^{th} *Grade* is a dummy for whether the students is in 8^{th} grade, and *Post* is a dummy for the year 2015. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Our coefficients are precisely estimated zeros, allowing us to reject positive effects larger than 0.008 SD in math, and 0.017 in language, well below the treatment effects found in the existing literature. In the teacher incentive program studied by Muralidharan and Sundararaman (2011) in India, average math and language test scores increased by 0.15 SD after one year, whereas Contreras and Rau (2012) find that a teacher incentive program in Chile had positive and large effects on language and math test scores of 0.14-0.25 SD. While the incentive scheme evaluated by Glewwe et al. (2010) in Kenya led to a 0.14 SD increase in test scores in tests linked to the incentive, the authors found no impact on the outcome of non-incentivized evaluations, consistent with our findings.

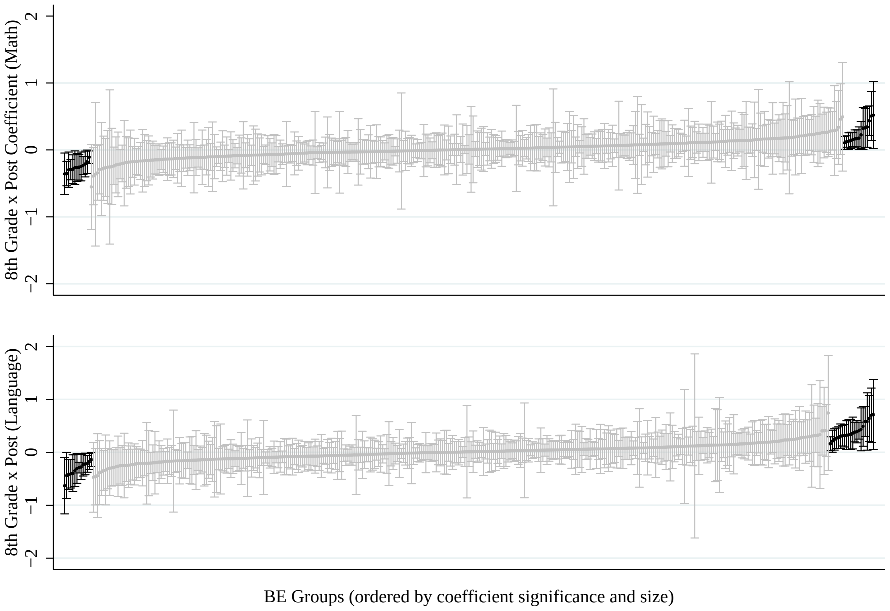
Since there are 395 distinct groups in which schools compete for the BE bonus, i.e., 395 different tournaments, we also evaluate the average effect of the teacher incentive in every competition. Figure 2.3 displays the 8^{th} *Grade* * *Post* coefficients (and its 95% confidence interval) for math and language in each of these 395 tournaments. In the vast majority of these groups, the teacher incentive had a zero average effect on student achievement. The coefficients for math and language are positive and statistically significant at the 5% level in only 4% and 6% of the BE groups,[30] providing further evidence of the BE's null average effect on student achievement.

As in most comparable studies, teacher bonuses under the BE are tied to students' performance in just two subjects (math and language). However, teacher incentives might also have an impact on student learning in other courses. The sign of this impact is theoretically unclear. On the one hand, schools could be tempted to devote more resources towards math and language at the expense of other subjects (e.g., augmenting instruction time), negatively impacting learning in the latter. On the other hand, 8th grade teachers in all subjects, not just math and language, might exert more effort knowing that their school's score largely rests on the performance of these students. Additionally, due to complementarities, if learning were higher in math and language, student achievement in

[30] Furthermore, in only 6 out of the 395 BE groups this holds simultaneously for math and language.

incentivized subjects might increase indirectly (Muralidharan and Sundararaman, 2011).[31] We do find a positive and small but significant effect of 0.011 SD on grades in non-incentivized courses, as shown in Table 2.4.

Figure 2.3: Effect of Teacher Incentive on Students' Math and Language Internal Grades in each BE Group



Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools which were eligible for taking the 2015 ECE standardized test and which are registered in the Ministry of Education's SIAGIE administrative system. The figures plot the 8^{th} *Grade* x *Post* coefficients and their 95% confidence intervals separately estimated for each BE group in math and language, respectively. BE groups in both figures are ordered by significance and coefficient size, and the ordering is separately done in each figure.

[31] Unlike studies carried out in primary school, math and language teachers are not responsible for teaching other subjects in secondary school. Thus, if there were any positive spillovers to other courses, they would be indirect.

Table 2.4: Effect of Teacher Incentive on Students' Internal Grades in Non-Incentivized Courses

	Non-Incentivized Courses		
	(1)	(2)	(3)
8^{th} Grade * Post	0.014***	0.014***	0.011***
	(0.004)	(0.004)	(0.004)
8^{th} Grade	-0.043***	-0.038***	-0.028***
	(0.003)	(0.003)	(0.002)
Repeated last year			-0.601***
			(0.006)
Male			0.284***
			(0.003)
Foreigner			0.047***
			(0.015)
Spanish is native tongue			0.120***
			(0.005)
Has a disability			-0.232***
			(0.012)
Father is alive			0.063***
			(0.003)
Mother is alive			0.044***
			(0.005)
Father lives in HH			0.022***
			(0.002)
Mother lives in HH			0.013***
			(0.002)
Observations	2108972	2108972	2108972
R^2	0.001	0.120	0.185
Year FE	YES	YES	YES
School FE	NO	YES	YES
Individual Controls	NO	NO	YES

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. The dependent variable is students' internal grades in non-incentivized courses, standardized by course-year. We first standardize each of the non-incentivized courses by course-year, and then take the average. Non-incentivized courses are art, science, social studies, English, civics, human relations, physical education, religion, and education for the workforce. 8^{th} *Grade* is a dummy for whether the students is in 8^{th} grade, and *Post* is a dummy for the year 2015. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table 2.A2 breaks the results down by each of the nine non-incentivized courses; we observe positive effects ranging between 0.014 SD and 0.017 SD in three cases (social studies, human relations, and religion). Although significant, the observed effect is very small, and well below the spillover effects found in other papers.[32] Furthermore, these results should be taken with caution because, as further discussed in Section 2.6.1, there is a divergence in the trend in non-incentivized courses in the year before the program was implemented.

2.5.1 Heterogeneous Effects

In a tournament such as the BE, if teachers are risk neutral, have symmetric information, and if students in all schools have the same ability (i..e, if all schools have the same ex-ante probability of winning), all teachers should exert the same effort as a result of the incentive, and who gets awarded the bonus should be random (Lazear and Rosen, 1981). However, if schools differ in their probability of winning, the incentive might not have the same power across the board. For example, teachers in schools in which students' pre-program levels of achievement are very far from the top 20% could be discouraged from exerting extra effort, and schools which are almost guaranteed to win might not respond to the incentive. This concern is partly mitigated in our setting by the fact that schools are grouped according to characteristics which are likely correlated with their students' performance, such as their number of hours of instruction, whether they are urban or rural and their school district. Nevertheless, important differences between schools in the same group might still remain, possibly affecting the reach of BE. This notion is brought forward in the Chilean study of Contreras and Rau (2012), where the authors find that the teacher incentive only had a positive impact on schools above the 65^{th} percentile in the distribution of pre-program score (the program awarded a bonus to schools in the top 25% within their group). The fact that the ECE was implemented in secondary schools for the first time in 2015 provides a limitation for performing this analysis in our context, since we cannot accurately determine a school's pre-tournament probability of winning.

[32] Muralidharan and Sundararaman (2011) find that teacher incentives targeted towards math and language standardized tests had an effect of 0.11 and 0.14 SD in science and social studies after only one year, an effect 10 to 13 times larger than the one we find.

As a second best, we proxy a school's likelihood of winning using its relative ranking within its BE group in terms of the socioeconomic status (SES) of its students. We construct an average measure of the SES of 8^{th} graders in 2015 by considering whether their first language is Spanish, and whether their parents have more than a primary school degree.[33] We then rank schools within their BE group according to this measure, and fully interact our baseline regression with 20 dummies indicating the percentile in the within BE group distribution that each school belongs to. As depicted in Figure 2.4, the estimates for all of these percentiles are very small in both math and language, and most of them are not statistically significant.[34] Having said this, it is highly unlikely that schools knew their relative standing in their BE group and could anticipate the likelihood of winning. We discuss this in Section 2.7.5.

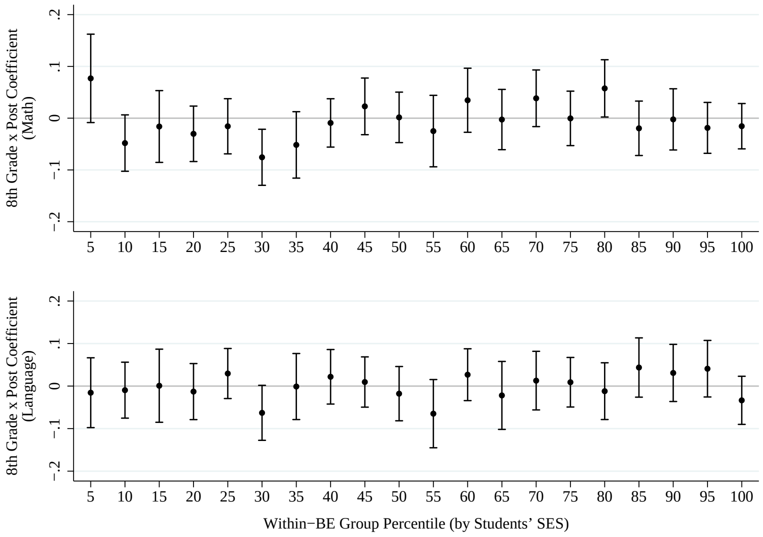
From a theoretical perspective, the strength of the incentive might be decreasing in the number of 8^{th} grade teachers and/or students, since the marginal impact of a teachers' effort on its school's score decreases when there are more teachers and students reached by the incentive, and teachers' ability to monitor each other also diminishes. For instance, Imberman and Lovenheim (2015) find that the effect of a group-based teacher incentive program in Houston is much stronger when teachers are responsible for teaching a higher share of students. Since our teacher database does not have information on the subject that each teacher is responsible for, we do not know how many incentivized teachers each school has; as a second best, we use the number of 8^{th} grade classes in 2015 as a proxy. We do not find any significant interaction of the BE incentive with enrollment or number of groups per grade, as seen in columns 6 and 7 of Table 2.5. Finally, we do not find any effects by whether the school ir urban or rural, as shown in column 8. As with any heterogeneity analysis, it is important to take the results with caution, since characteristics such as

[33] For each 8^{th} grader in 2015, we add three dummy variables: whether his first language is Spanish, and dummies for whether his mother and father have more than a primary school education. We then calculate the average index for each school.

[34] We also perform this exercise ranking schools instead by an index measuring the quality of their infrastructure, another proxy of their probability of winning, and find no discernible patterns either (results upon request).

enrollment and urbanicity are not randomly assigned, and could be proxying for something else. Ideally, we would also be able to test for heterogeneous effects across teacher characteristics. Unfortunately, although we know who the teachers are for each class, we do not know which of the teachers teach math and language.

Figure 2.4: Heterogeneous Effect of Teacher Incentive on Students' Internal Grades by Schools' Socioeconomic Status (SES) Rank



Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. The dependent variables in the top and bottom figures are students' internal grades in math and language, respectively, standardized by course-year. We construct a SES index (taking values 0-3) for each 8^{th} student in 2015, adding up three dummies for whether his first language is Spanish, and whether his mother and father have more than a primary school education. We then calculate the average index for each school, and group schools in 20 percentile gruops according to their ranking in terms of this measure within their BE group. The figures plot the coeffient and 95% confidence interval for the interaction of each percentile dummy and the 8^{th} * Post dummy from our baseline regressions.

Following other papers in the literature, we also test for heterogeneous effects across gender, by whether students' first language is Spanish, and by their parents' educational attainment. The latter variable is an index from 0 to 2, taking a value of 0 if both parents have a primary school degree or less, 1 if one parent has more than primary schooling, and 2 if both do. Parents' education and students' native tongue are proxies for socioeconomic status in Peru. As displayed in column 1 of Table 2.5, and consistent with the finding in Muralidharan and Sundararaman (2011) and Behrman et al. (2015), we do not find any heterogeneity by gender. Neither do we find heterogeneous effects by socioeconomic status, proxied by native language and parents' education (columns 2 and 3). The literature is mixed on this particular issue, since Muralidharan and Sundararaman (2011) observe that students from more affluent families have a stronger response to the teacher incentive program, whereas Lavy (2002) finds that it is students with poor socioeconomic backgrounds that benefit more from it. However, the program evaluated in the latter was designed so as to encourage teachers to focus on weak students.

Considering that teachers might focus on certain students, and student responsiveness might vary according to prior achievement, we also test for heterogeneity across measures of students' past performance, namely whether the student was retained in the previous year and by the student's lagged internal grade in the same subject (standardized by school, grade and year).[35] Pay-for-performance programs in which bonus payments depend on whether students attain a certain threshold, such as passing an exam, create incentives for teachers to focus on students close to this cutoff (e.g. Lavy, 2009 and Neal and Schanzenbach, 2010). On the contrary, if obtaining the bonus depends on the average score, such as in the BE program under analysis, teachers will find it optimal to target students most responsive to any increased teacher effort. If the function mapping teacher effort into test score gains is concave (convex) in past performance, teachers would react by focusing more intensely on the weaker (stronger) students (Muralidharan and Sundararaman, 2011). However, as shown in columns 4 and 5, we do not find any heterogeneity according to students' past performance.[36] These results are consistent with the findings of Behrman et al. (2015).

[35] Lagged grades are only available for students in 2014 and 2015, since our database only has student identifiers which can be linked across years starting 2013. Importantly, if we restrict our sample to this period, results on average treatment effects do not change.

[36] We also perform this estimation by grouping students into quintiles and terciles of the distribution of lagged grades in their same school, grade and year. The results are unchanged, as reported in Table 2.A3.

Table 2.5: Heterogeneous Effect of Teacher Incentive on Students' Internal Grades

		Spanish	Parents		Lagged		Num.	
	Male	Speaker	High Educ	Repeated	Grade	Ln Enrollment	Classes	Rural
Panel A: Math Grades								
8^{th} Grade * Post	-0.010	-0.004	-0.005	-0.003	-0.002	0.033	-0.002	-0.007
	(0.008)	(0.012)	(0.007)	(0.007)	(0.008)	(0.023)	(0.009)	(0.008)
8^{th} Grade * Post * Covariate	0.010	-0.000	-0.003	-0.019	0.006	-0.008	-0.001	0.016
	(0.009)	(0.013)	(0.005)	(0.016)	(0.005)	(0.006)	(0.002)	(0.012)
Observations	2108806	2108806	1851727	2108806	1382813	2108806	2108806	2108806
R^2	0.092	0.092	0.099	0.092	0.440	0.092	0.092	0.092
Panel B: Language Grades								
8^{th} Grade * Post	0.006	-0.000	-0.004	0.002	0.005	-0.023	-0.008	0.004
	(0.009)	(0.013)	(0.009)	(0.008)	(0.009)	(0.027)	(0.010)	(0.009)
8^{th} Grade * Post * Covariate	-0.010	0.002	0.003	-0.008	0.001	0.006	0.002	-0.017
	(0.010)	(0.015)	(0.006)	(0.016)	(0.005)	(0.007)	(0.002)	(0.014)
Observations	2108793	2108793	1851715	2108793	1382756	2108793	2108793	2108793
R^2	0.135	0.135	0.142	0.135	0.417	0.135	0.135	0.135

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. Heterogeneities by *Lagged Grade* exclude the year 2013 for which students' previous grade is unavailable, and heterogeneities by parents' education exclude 12% of students in 2013-2015 for which this variable is missing. The dependent variables are students' internal grades in math and language, standardized by course-year. 8^{th} *Grade* is a dummy for whether the students is in 8^{th} grade, *Post* is a dummy for the year 2015, and *Covariate* is the variable indicated in the column header. All regressions include school and year fixed effects, as well as the standard individual controls and the three-way interaction between 8^{th} *Grade*, *Post* and *Covariate*. We only report two coefficients for exposition purposes. *Spanish Speaker* is a dummy for whether the student's first language is Spanish, and *Parents High Educ* is 0 if both parents have a primary school degree or less, is 1 if only one of the parents has more than a primary school degree, and 2 if both. *Repeated* is a dummy for whether the student was retained in the same grade at the end of the previous year, and *Lagged Grade* is the students' internal grade in that particular course in the previous year, standardized by school and grade. *Ln Enrollment* is the log of the number of students enrolled in that year and grade. *Num. Classes* is the number of classes in the student's grade and year, *Rural* is a dummy for whether the school is in a rural area. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

2.6 Testing the Validity of the Identification Strategy

This section provides further evidence on the validity of our difference-in-differences estimation strategy. We provide formal evidence in support of the parallel trends assumption, and demonstrate that internal grades are broadly correlated with ECE test scores, and vary considerably within schools. Furthermore, we corroborate that our null effects are not driven by positive spillovers to our comparison school, and show that schools did not change the way in which they assigned teachers across grades as a result of the teacher incentive program.

2.6.1 Parallel Trends

To test whether there is a divergence in the trends of 8^{th} and 9^{th} grade students in 2014, we add an interaction between the 8^{th} grade dummy and an indicator for 2014 to our baseline specification. Reassuringly, the coefficients for the pre-treatment difference-in-differences are precisely estimated zeroes for both math and language, as shown in Table 2.6. In the case of non-incentivized courses, however, there is a relative increase in 8^{th} graders' internal grades in 2014. Although the magnitude of this change is small (0.010 SD), it is similar in magnitude to the estimated impacts for 2015. Hence, the results using non-incentivized courses as an outcome should be taken with caution.

2.6.2 Internal Grades Reflect Learning

Unlike other studies on teacher pay-for-performance, we measure learning using students' internal grades instead of their standardized test results.[37] As discussed in Section 2.3, internal grades have the advantage of capturing student achievement without directly influencing teachers' probability of obtaining the bonus. However, internal grades are subjectively assigned by teachers, and are not awarded using a uniform criterion as standardized tests are. Since each school

[37] In a recent study, Chong et al. (2016) also use internal grades to measure student achievement in rural Peru.

Table 2.6: Test for Parallel Trends in Students' Internal Grades

	Math	Language	Non-Incentivized Courses
8^{th} Grade * Post	-0.004	0.002	0.016***
	(0.007)	(0.009)	(0.005)
8^{th} Grade * 2014	0.002	0.001	0.010**
	(0.007)	(0.009)	(0.004)
8^{th} Grade	-0.009*	-0.017***	-0.033***
	(0.005)	(0.007)	(0.003)
Observations	2108806	2108793	2108972
R^2	0.092	0.135	0.185

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE are registered in the Ministry of Education's SIAGIE administrative system. All regressions include year fixed effects, school fixed effects, and the standard controls. The dependent variables are students' internal grades in math, language and non-incentivized courses, standardized by course-year. We take the average of non-incentivized courses, which are art, science, social studies, English, civics, human relations, physical education, religion, and education for the workforce. 8^{th} *Grade* is a dummy for whether the students is in 8^{th} grade, *Post* is a dummy for the year 2015, and *2014* is a dummy for the year 2014. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

might have its own grading standards, making differences in internal grades not necessarily reflective of differences in learning across schools, we restrict our comparison to students from the same school to control for school-specific grading standards.[38] What is crucial for identifying a causal effect is that internal grades capture changes in learning across different grades within the same school. That is, if 8^{th} grade students in a particular school are learning more as a result of the teacher incentive, the relative internal grades of 8^{th} graders

[38] Although it would be preferable to include teacher fixed effects to control for teachers' grading standards, we only know the grades and classes teachers are assigned to, but not the subject that they teach. We cannot identify who the teacher handing out the grades for each subject is, and therefore cannot include teacher fixed effects in our estimation. However, since teachers are not systematically changing across 8^{th} and 9^{th} grades, as shown in Section 2.6.3 below, unobserved teacher characteristics are unlikely to bias our estimates.

in that school should rise. We face two potential threats in this regard. Firstly, teachers might not award internal grades in a systematic way. This doubt is raised by the findings of a few papers comparing grading standards in blind versus non-blind examinations. While some studies find evidence of discrimination in grading based on students' gender (Lavy, 2008), ethnicity (Botelho et al., 2015; Burgess and Greaves, 2013), and caste (Hanna and Linden, 2012), others find no such disparities (Newstead and Dennis, 1990; Baird, 1998; Van Ewijk, 2011). Consistent with the latter, we show that student characteristics correlate with internal grades and with standardized test scores in a consistent manner within the schools in our sample, alleviating this concern. A second threat to our identification is that if teachers grade on a relative basis (e.g., the worst 10% always fails, or the top 10% always gets the highest grade), we might not be able to detect overall changes in student learning using internal grades. It turns out, however, that there is substantial variation in the distribution of grades across classes and years in the same school.

Considering that our identification requires that internal grades reflect within-school differences in learning, standardized test scores and internal grades should broadly follow the same patterns when comparing students from the same school. Unfortunately, ECE test scores are disclosed at the school level, meaning that for every secondary school taking the ECE in 2015, we only observe the mean score in math and language, as well as the fraction of 8^{th} graders with very low, low, medium and high performance. Given our data limitations, we examine the cross-sectional correlation between average ECE scores and the average internal grades of 8^{th} graders in 2015, for the public secondary schools in our sample.[39] We also explore the correlation between the fraction of students who fail math and language according to their internal grades, and the fraction of low performing students in the ECE. To facilitate the interpretation of the coefficients, we express average internal grades and average ECE scores as a z-score, and control for school district fixed effects, school characteristics and the average

[39] Our sample for this analysis (8,010 schools) is slightly smaller than our baseline sample of 8,059 schools because a few schools with were eligible to take the ECE (and were thus eligible to participate in the BE) ended up not taking the test, or were faced with problems during its implementation.

characteristics of students from each school. As shown in Appendix Table 2.A4, average ECE scores and internal grades are positively and significantly correlated, although their correspondence is relatively weak. In particular, a 1 standard deviation increase in average math (language) internal grades is associated with an increase in average ECE scores of 0.116 (0.103) SD. Moreover, a 1 percentage point increase in the share of students failing math (language) according to their internal grades corresponds to a 0.071 (0.091) percentage point rise in the proportion of students with the lowest attainment in the ECE.[40,41]

Having said this, it is hard to establish whether internal grades reflect learning by just comparing the aggregate cross-sectional correlation of these and ECE grades. For one, internal grades might capture a related but different dimension of learning than standardized test scores. Additionally, since internal grades are likely to depend on school grading standards, it is unclear that they can be compared across schools.[42] While the disclosure of ECE test scores does not allow us to identify students' individual performance, the Peruvian Ministry of Education provides an anonymized database with individual ECE test scores, gender, an index of socioeconomic status (constructed using parents' education, and household assets and characteristics), and anonymized school identifiers. As shown in Panel A of Table 2.A6, students are more likely to obtain a higher ECE test score in math and language if they are male and have a high socioeconomic status, as compared to other students from the same school. An analogous

[40] In 2015, 55% and 43% of 8^{th} graders in the average school were ranked in the lowest category according to their ECE scores, whereas the average school only had 13% and 9% of their 8^{th} graders failing math and language, respectively. These two categorizations are only broadly comparable, and these results must thus be taken with caution.

[41] We also examine whether school and average student characteristics explain internal and ECE grades in a similar manner, by separately regressing schools' 2015 average ECE and internal grades against a series of controls. As displayed in Appendix Table 2.A5, the same broad patterns hold for both types of grades in math and language. Schools in which a high proportion of students have parents with more than a primary school degree do better as reflected by both ECE and internal grades. The same holds for schools with longer school days, and schools in which a high proportion of students have Spanish as their first language, Furthermore, schools in which a higher share of 8^{th} graders were retained the year before do worse according to both measures.

[42] If good schools set harsh grading standards, and low quality schools are lenient in their grading, for example, differences in the average internal grades of these two types of schools will not convey any information on their differences in student achievement.

regression with 8^{th} grade students' individual internal grades as the dependent variable (Panel B of Table 2.A6) shows that the within-school correlation between student achievement and gender and socioeconomic status is qualitatively similar. Despite the fact that internal grades and standardized test scores are prone to measure learning differently, and that students have different stakes in each of these outcomes, these two measures seem to relate in a consistent manner when comparing students from the same school.

Having established that internal grades are correlated with standardized test scores, we now provide evidence of the fact that grading on a curve is uncommon in Peruvian secondary schools. If teachers were assigning grades on a relative basis, we would expect two different classes in the same school, grade and year to have a very similar grade distribution. Our database on teachers shows that on average, 8^{th} grade teachers from schools with only two classes teach in 92% of them, meaning that the teachers handing out the grades are practically the same across classes. We restrict our sample to 8^{th} graders in schools with just two 8^{th} grade groups in 2014 (accounting for 17% of our schools), and test whether math and language internal grades have a different mean and standard deviation across both classes belonging to the same school. With a significance level of 10%, in 23% and 32% of cases we reject the null hypothesis of equal means across both groups in math and language, respectively. The average difference in means across groups is 0.66 and 0.77 in math and language, roughly one third of a standard deviation. An F-test for the equality of variances shows that in 23% and 21% of our schools, we can reject the null hypothesis that the distribution of math and language grades has the same standard deviation.[43] The difference in means and standard deviations and their corresponding p-values are depicted in Figure 2.A1. All in all, this evidence points to the fact that grading on a curve is not the norm in Peruvian high schools.

[43] The average difference in standard deviations is 0.44 in math and 0.41 in language.

2.6.3 No Spillovers to 9^{th} Grade Students

A crucial condition for identifying the causal impact of BE is that the performance of 9^{th} grade students, our comparison group, should be unaffected by the program. One possible concern is that the incentives introduced by BE could lead schools to change the assignment of teachers across 8^{th} and 9^{th} grade in 2015. If schools assigned the best teachers to 8^{th} grade in 2015, for instance, our estimates would be upward biased. However, since we find that BE had a null effect on student achievement, this is less of a concern. Furthermore, since the program was only announced after the school year had started, as illustrated in Figure 2.1, it would have been hard for schools to shift teachers around. We still perform some tests to learn whether schools teacher characteristics changed differently across grades in the year in which BE was introduced. Given that we cannot identify which teachers are responsible for instructing math and language, and do not have an objective measure of teacher quality, it is hard to test if BE brought about changes in the average quality of teachers across grades. However, we do observe the school, grades and classes to which teachers are assigned to in 2013-2015, and have some observable teacher characteristics which might be correlated with their performance. We use this information to test for differential changes in 2015 in the average characteristics of 8^{th} and 9^{th} grade teachers from the same school. As shown in Table 2.7, we do not find any differential change in the average age and gender of teachers in 2015. Neither do we observe a significant change in the proportion of teachers who are new to the school or new to that particular grade and school, or in the average number of courses taught by teachers in that grade. We do observe a significant decrease in the average number of secondary schools in which 8^{th} grade teachers are working. Although this might mean that 8^{th} grade teachers were less time constrained in 2015, this only represents a 0.3$\%$ drop from the mean. Consequently, there is no strong evidence of changes in teacher composition across 8^{th} and 9^{th} grade classes belonging to the same school in 2015.

Table 2.7: Test for Changes in Teacher Composition Across Grades

	Average Age	% Male	Average Number of Classes	Average Number of Schools	% New to School	% New to School-Grade
8^{th} Grade * Post	-0.013	0.001	0.016	-0.005**	-0.000	-0.000
	(0.021)	(0.001)	(0.016)	(0.002)	(0.001)	(0.002)
8^{th} Grade	0.043***	-0.006***	-0.090***	0.000	0.001	0.001
	(0.013)	(0.001)	(0.011)	(0.001)	(0.001)	(0.001)
Observations	46614	46615	46615	46615	31332	31332
R^2	0.869	0.738	0.953	0.737	0.808	0.758
Mean Dep. Variable	41.685	0.598	11.790	1.634	0.474	0.552

Notes: The sample includes all public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE standardized test, registered in the Ministry of Education's SIAGIE administrative system, and with data on teacher characteristics. The unit of analysis in these regressions is a school-grade-year, for 8^{th} and 9^{th} grade. *Average Age* is the average age of teachers in that grade, and *% Male* is the % of teachers in that grade that are male. *Average Number of Classes* is the average number of courses taught by teachers in that grade, and *Average Number of Schools* is the mean number of different secondary schools in which the teacher works. *% New to School-Grade* are the proportion of teachers in that particular grade who are new to the school, or new to that particular grade , respectively. All regressions include school fixed effects, year fixed effects, a dummy for 8^{th} *Grade*, and the interaction between 8^{th} *Grade* and a dummy for 2015 (i.e., *Post*). The regressions in columns 5 and 6 do not include the year 2013 since we do not have information on teachers' appointments in 2012. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

A bigger concern given our findings is that BE improved teacher behavior overall, instead of impact teaching to 8^{th} graders differently. As we explain in Section 2.2.2, in practice around 80% of a school's score depends on the performance of 8^{th} grade students in the ECE standardized tests. Thus, a very small portion of the school's score could be improved if 9^{th} grade teachers exerted more effort. However, since 83% of 8^{th} grade teachers also instruct 9^{th} grade, any increase in effort while teaching 8^{th} graders could spill over to students in our comparison group, biasing our estimation downwards. Alleviating this concern, we find that the effects are also null in schools in which a low share of 8^{th} grade teachers also instructs 9^{th} graders (columns 1, 2, 4 and 5 of Table 2.8).[44] If anything, there is a significant (though very small) positive effect in math grades in schools in which most teachers instruct both 8^{th} and 9^{th} graders (column 3).

2.7 Why Didn't Student Learning Increase?

Having established that student learning did not increase as a result of the teacher incentive program, this section discusses and provides suggestive evidence on a series of potential explanations for why the program had a null effect.

2.7.1 Teachers Did Not Know About the Program or Did Not Understand It

An explanation for the null effects we find is that schools simply did not hear about the BE, or did not understand the formula by which scores were calculated. We argue that this is at best a partial explanation. In 2015, along with the launch of the 2015 edition of the BE program, the Ministry of Education headed a diffusion campaign, making it likely that secondary school teachers were informed about the program. Furthermore, the fact that the principal and every teacher get paid if their school wins generates strong incentives for people working in the same institution to inform each other about the BE. In our teacher survey, 64%

[44] Even though what matters is the overlap of math and language teachers, we do not have information on the subjects taught by each teacher, and are thus restricted to perform this analysis using the average overlap of all teachers across 8^{th} and 9^{th} grade.

Table 2.8: HETEROGENEITY BY OVERLAP OF 8^{th} AND 9^{th} GRADE TEACHERS

	Math			Language		
	Low	Med	High	Low	Med	High
8^{th} Grade * Post	-0.014	-0.014	0.013*	0.010	0.001	-0.009
	(0.014)	(0.012)	(0.007)	(0.017)	(0.014)	(0.008)
8^{th} Grade	-0.001	-0.006	-0.016***	-0.017	-0.017**	-0.014***
	(0.008)	(0.007)	(0.004)	(0.011)	(0.008)	(0.005)
Repeated last year	-0.594***	-0.533***	-0.582***	-0.637***	-0.587***	-0.642***
	(0.013)	(0.008)	(0.009)	(0.014)	(0.009)	(0.010)
Male	0.123***	0.128***	0.097***	0.369***	0.367***	0.275***
	(0.006)	(0.005)	(0.004)	(0.007)	(0.005)	(0.005)
Foreigner	0.048*	0.017	0.124**	0.084***	0.082***	0.142**
	(0.027)	(0.028)	(0.055)	(0.027)	(0.029)	(0.057)
Spanish is native tongue	0.130***	0.180***	0.198***	0.129***	0.185***	0.202***
	(0.013)	(0.012)	(0.009)	(0.013)	(0.013)	(0.009)
Has a disability	-0.231***	-0.247***	-0.300***	-0.200***	-0.243***	-0.310***
	(0.028)	(0.025)	(0.019)	(0.026)	(0.027)	(0.021)
Father is alive	0.070***	0.068***	0.061***	0.073***	0.061***	0.065***
	(0.007)	(0.006)	(0.005)	(0.006)	(0.006)	(0.005)
Mother is alive	0.020	0.044***	0.055***	0.029**	0.051***	0.067***
	(0.013)	(0.010)	(0.008)	(0.011)	(0.010)	(0.009)
Father lives in HH	0.034***	0.032***	0.011***	0.028***	0.023***	0.006
	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Mother lives in HH	0.013***	0.012***	0.020***	0.015***	0.017***	0.025***
	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	740453	672268	696085	740446	672275	696072
R^2	0.072	0.084	0.126	0.116	0.129	0.166
Avg. % of teachers in both grades	0.392	0.654	0.906	0.392	0.654	0.906

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. Columns Low, Med and High restrict the sample to students in schools with a low, medium and high average overlap between 8^{th} and 9^{th} grade teachers in 2013-2015. Overlap between 8^{th} and 9^{th} grade teachers is the % of teachers in 8^{th} grade also instructing in 9^{th} grade. The dependent variables are students' internal grades in math and language, standardized by course-year. 8^{th} *Grade* is a dummy for whether the students is in 8^{th} grade, and *Post* is a dummy for the year 2015. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

of those who taught math or language in 8^{th} grade in 2015 reported that they knew about the program's existence during the 2015 academic year. When asked about how they heard about BE, 57% answered that they found out through the Ministry of Education or the school district authorities, 30% answered that they got the information from the news, and 35% from the school principal or other coworkers (they could select more than one option).

Although we only evaluate the effect of BE in its first year, the system by which schools were scored under the BE was not overly complex.[45] It should have been relatively clear from a teacher's perspective that the main component of his/her school's score is the average performance of 8^{th} graders in the ECE standardized tests. This stems from the fact that performance in the ECE test was the main component of schools' scores in the two previous rounds of the BE in primary schools. The BE program had already been going on for two editions in every public primary school in the country, and the experience of primary schools with the BE was salient in the national news.[46] This is broadly confirmed by our survey, in which 64% of math or language 8^{th} grade teachers who knew about BE in 2015 answered that ECE test scores were the most important or second most important component of the BE score.

Almost half of the schools in our sample share the building with a primary school that participated in the BE before, and 13% of them operate in the same building as primary school BE winner. Even though the salience of the secondary school BE was probably higher in these cases, we do not find any effects on math and language test scores in either of these groups of schools, as shown in columns (3), (4), (8) and (9) of Table 2.9. Thus, it is unlikely that the BE had no impact because of schools' lack of awareness of its existence or its rules.

[45] Other studies on teaching incentives with similar formulas for assigning the bonus (Lavy, 2002 and Contreras and Rau, 2012) find positive and significant effects on student learning.

[46] For instance, http://larepublica.pe/23-10-2014/maestros-tendran-bono-de-hasta-3-mil-soles-por-buen-desempeno and http://www.andina.com.pe/agencia/noticia-bono-hasta-s-3000-buen-desempeno-docente-se-pagara-noviembre-528482.aspx

Table 2.9: Effect of Teacher Incentive by Average Salary, Number of Classes, School's Experience with Primary School BE and BE Group Size

	Math					Language				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
8^{th} Grade * Post	0.249	-0.011	-0.006	-0.007	-0.014*	0.299	0.000	0.005	-0.000	-0.010
	(0.501)	(0.008)	(0.009)	(0.007)	(0.008)	(0.585)	(0.010)	(0.011)	(0.009)	(0.011)
8^{th} Grade * Post * Ln (Average Salary)	-0.035					-0.041				
	(0.069)					(0.081)				
8^{th} Grade * Post * One Class		0.030***					0.004			
		(0.011)					(0.013)			
8^{th} Grade * Post * BE Primary			0.003					-0.007		
			(0.013)					(0.016)		
8^{th} Grade * Post * BE Primary Winner				0.016					0.010	
				(0.018)					(0.024)	
8^{th} Grade * Post * Small BE Group					0.026**					0.033**
					(0.013)					(0.016)
Observations	2077227	2108806	2108806	2108806	2108806	2077214	2108793	2108793	2108793	2108793
R^2	0.092	0.092	0.092	0.092	0.092	0.135	0.135	0.135	0.135	0.135
P-Value (sum of both coefficients)	0.621	0.008	0.739	0.568	0.237	0.601	0.558	0.893	0.669	0.062

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. The dependent variables are students' internal grades in math and language, standardized by course-year. 8^{th} *Grade* is a dummy for whether the students is in 8^{th} grade, *Post* is a dummy for the year 2015, *Ln (Average Salary)* is the average salary of teachers in each school in 2015 (in logs), obtained from the 2015 school census, and *One Class* is a dummy for whether the student attends a school in which there is only one class in his grade. *BE Primary* is a dummy for whether a primary school that participated in the BE in the past operates in the same building, and *BE Primary* is a dummy for whether there is a primary school in the building that won the BE bonus in the past. *Small BE Group* is a dummy for whether the number of schools in the corresponding BE group is below the median. All regressions include year and school fixed effects, the standard individual controls, and the three-way interaction between 8^{th} *Grade*, *Post* and the specific heterogeneity variable. We only report two coefficients for exposition purposes. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

2.7.2 The Incentive Was Too Small

The prize that teachers could receive under the BE is in the range of bonuses granted in other studies finding positive effects. As described in Section 2.2, the BE bonus corresponds to either 1 or 1.4 monthly salaries of the average teacher, and was awarded to teachers in 20% of schools. The average bonus represents approximately 24% of a monthly salary, making this incentive sizable in comparison to that of other studies, in which the average value of the prize ranges between 3% and 35% of a monthly salary.[47] In the subsample of 8^{th} grade math or language teachers who responded our survey, 42% correctly identified the bonus amount or thought that it was larger, 20% did not know the exact bonus amount, 2% thought that it was smaller, and 36% did not know about the BE in 2015. However, when we asked their opinion on the size of the prize, only 30% of those who knew about the program thought that the prize was adequate or large. This may have to do with the fact that the survey was coming from the Ministry of Education, and many teachers took this as an opportunity to complain about their low salaries.[48]

If the bonus were not large enough to incentivize the average teacher, we would perhaps find a positive effect in schools in which teachers' pay is relatively low. However, as shown in columns (1) and (6) of Table 2.9, we do not find any heterogeneity by teachers' average salary in 2015.[49] Although we cannot exclude that the incentive scheme would have worked with a larger bonus, there is no evidence that the size of the incentive is the reason why the program had no

[47] The average bonus obtained by Indian teachers in Muralidharan and Sundararaman (2011) is around 35% of a monthly salary, whereas bonuses in the experiment run by Glewwe et al. (2010) in Kenya have an average value ranging between 12% and 21% of a teachers' monthly wage. The incentive implemented in Chile and studied by Contreras and Rau (2012) paid teachers 10% of a monthly salary on average. Finally, the Israeli program studied by Lavy (2002) awards prizes of 10%-40% of an average teacher's monthly salary to approximately one third of participating teachers.

[48] In the open-ended part of this question, many teachers answered that their salaries are insufficient. Furthermore, quite a few teachers answered the survey email with complaints about their working conditions.

[49] We calculate the average salary of teachers in every secondary school from the number of contract teachers and civil servant teachers in each pay scale, as reported in the 2015 school census. Since the school census does not provide disaggregated data by grade, we take the school average.

distinguishable effect on students' math and language grades.

2.7.3 Group Incentives Do Not Work

When incentives are collective, the mapping of a teachers' actions on his/her probability of obtaining the bonus is weaker when the number of teachers reached by the incentive is larger, raising the likelihood of free-riding (Holmstrom, 1982), and thus lowering the incentive's power in promoting higher teacher effort. While collective incentives have the potential of inducing higher cooperation and monitoring among teachers (Kandel and Lazear, 1992; Kandori, 1992), this might be harder to achieve when the number of incentivized teachers is very large. Although we do not know the fraction of 8^{th} grade students that each math and language teacher instructs (we do not have information on the subject taught by teachers), we do know the number of 8^{th} grade classes that each school has in 2015. In 2013-2015, the average secondary school in our sample had only two groups of 8^{th} graders. Since there is at most one math and language teacher per group, the average school has no more than four incentivized (i.e., math and language) teachers, a figure comparable to the number of incentivized teachers in other papers in the literature finding positive effects when teacher incentives are collective. The average school in Muralidharan and Sundararaman (2011) and Glewwe et al. (2010) has three and six incentivized teachers, for example.

As shown in column (7) of Table 2.5, we don't find any differential effects by the number of 8^{th} grade groups. If we break the results down even more, as shown in column (2) of Table 2.9, we do find that the BE had a small but significant positive effect in the math grades of students in schools with only one class per grade (accounting for 21% of students in 61% of schools). More specifically, the teacher incentive increases math grades by 0.019 SD,[50] although these effects are much smaller than those found in the other studies in the literature. Thus, the fact that the incentive faced by teachers under the BE is collective does not seem to be one of the main reasons why the program had no effect, although it might have

[50] The sum of the 8^{th} *Grade* $\times$ *Post* and 8^{th} *Grade* $\times$ *Post* $\times$ *One Class* coefficients yields a total effect of 0.019 SD, with a p-value of 0.008.

some bite.

2.7.4 Teachers Only Focused on Improving Standardized Test Scores

As discussed in Section 2.3, teacher incentive programs might not result in higher learning if teachers focus their efforts on short-term strategies aimed solely at increasing standardized test scores. Teachers might have reacted to the incentive by targeting topics likely to appear in the ECE, coaching students on test-taking strategies, or even cheating. Since 2015 was the first year in which students took the ECE, and there is no appropriate control group (every public school in which 8^{th} graders participated in the ECE is also eligible for the BE), we cannot identify whether ECE test scores increased as a result of the teacher incentive program. Thus, we cannot initially rule out this hypothesis. However, there are reasons why we believe that teachers could not engage in this type of behavior. Firstly, independent officials, trained and working directly for the Peruvian National Statistics Institute were in charge of the implementation of the ECE. Teachers were not allowed to be in the room at any moment during the exam and were not responsible for its correction. Thus, it is very unlikely that schools could cheat.[51] Secondly, because the ECE exam is designed to capture a wide range of skills,[52] teachers could hardly influence this outcome by narrowing their instructional focus on certain topics. Thirdly, due to the fact that the ECE was implemented for the first time in secondary schools in 2015, secondary school teachers did not have previous experience with this type of standardized tests and, consequently, could hardly predict the content or the specific format of the exam. As the content of the standardized exam was not predictable, coaching or narrow teaching are less of a concern in this setting (Neal, 2011).

Having said this, our online survey inquired about whether teachers changed

[51] Since students had no personal stake in this exam, there were no incentive to cheat on their part.

[52] Details on the design of the ECE are reported by the Ministry of Education in *Reporte Técnico de la Evaluación Censal de Estudiantes (ECE 2015)*, available at http://umc.minedu.gob.pe/wp-content/uploads/2016/07/Reporte-Tecnico-ECE-2015.pdf.

their pedagogical practices in 2015 as a result of BE, and separately asks about their pedagogical changes while teaching 8^{th} grade as opposed to all other grades. Table 2.10 reports the results of this question for all math and language teachers taking the survey who reported that they knew about BE in 2015 (those who did not know where not asked this question). These results must be taken with caution, since it is probable that there was some bias in reporting given the framing of the survey in terms of the BE program.[53] As can be seen in Panel A, 8^{th} grade teachers are 5 percentage points more likely to report that they improved their attendance, and 10 percentage points more likely to report that they gave their students more homework, evaluated them more often and/or gave extra tutoring sessions, as compared to math and language teachers from other grades. There are statistically significant differences as well in how often they report that they paid attention to the weakest students (5 percentage points), increased the difficulty of their classes (6 points), and increased the frequency of multiple choice examinations (9 percentage points). The same patterns hold when we restrict the analysis to teachers that taught math or language in 8^{th} grade and other grades, as seen in Panel B. While some of these self-reported differences in teacher behavior are consistent with teaching-to-the-test (e.g., increasing the frequency of multiple choice evaluations), if teachers were in fact improving their attendance or paying more attention to the weakest students, student achievement in terms of internal grades should have increased, and it did not.

[53] In the study of Glewwe et al. (2010), for example, the survey to teachers was also framed as soliciting feedback on the incentive program; teachers in the treatment group were more likely to report having increased the number of homework assignments, whereas student reports suggest no such differences. In Behrman et al. (2015), teachers were also more likely to report that they spent more hours preparing their students for the test, although the incentive had no impact on student outcomes.

Table 2.10: Effect of Teacher Incentive on Teachers' Pedagogical Practices

	8^{th} Grade	Other Grades	Difference	P-Value
Panel A: All Math/Language Teachers				
Improved attendance	0.207	0.157	0.050**	0.024
More homework, evaluations and/or tutoring sessions	0.471	0.370	0.101***	0.000
Paid more attention to weakest students	0.683	0.637	0.046*	0.097
Training programs or feedback sessions	0.548	0.542	0.006	0.828
Increased difficulty of classes	0.192	0.135	0.056***	0.007
Decreased difficulty of classes	0.148	0.138	0.010	0.620
More multiple choice tests	0.385	0.299	0.086***	0.002
Other	0.130	0.150	-0.020	0.317
Number of teachers	454	865		
Panel B: Math/Language Teachers in Both Grades				
Improved attendance	0.203	0.143	0.060***	0.000
More homework, evaluations and/or tutoring sessions	0.460	0.326	0.134***	0.000
Paid more attention to weakest students	0.677	0.657	0.020	0.209
Training programs or feedback sessions	0.523	0.494	0.029	0.149
Increased difficulty of classes	0.197	0.157	0.040**	0.016
Decreased difficulty of classes	0.146	0.131	0.014	0.298
More multiple choice tests	0.391	0.337	0.054**	0.017
Other	0.123	0.143	-0.020	0.250
Number of teachers	350	350		

Notes: The sample includes all survey respondents who taught math or language in 8^{th} and other grades in 2015, and knew about the BE program during the 2015 academic year. Panel B only includes those who taught math or language in 8^{th} grade and other grades. Teachers were asked whether they changed their pedagogical practices in 2015 as a result of BE, and could answer any of the options specified in the table rows. We asked them separately about changes while teaching 8^{th} grade (column 1) as opposed to all other grades (column 2), in case the teacher taught both. * significant at 10%; ** significant at 5%; *** significant at 1%.

2.7.5 Teachers Were Unfamiliar with the Standardized Test and Students Had No Stakes in It

Given that 2015 was the first year in which the ECE test was implemented in secondary schools, teachers might have been uncertain about the function mapping their effort into students' ECE test scores. The connection between teachers' effort and their expected benefit might have therefore been diluted, making the incentive insufficient for prompting teachers into exerting more effort (Fryer, 2013).[54] Even if teachers knew how to equip their students with the skills needed to obtain high ECE scores, they might have encountered difficulties in passing on the incentive to their students. Since ECE tests have no impact whatsoever on students' grades, and the Ministry of Education only reports school averages (and not individual test scores) to schools, teachers, parents and even students, the latter might have little or no incentive to put effort in these tests.[55] Teachers might have anticipated that their actions would only marginally impact their students' ECE scores, and thus might have been discouraged from exerting more effort. The results from the experimental study implemented by Behrman et al. (2015) in Mexico provide suggestive evidence on the possibility that incentivizing teachers on their students' performance might not be effective unless students have a stake as well.[56] This hypothesis is partially supported by

[54] A series of experimental studies in rural India suggest that teachers' knowledge and incentives might be complementary inputs in the education production function. While Muralidharan and Sundararaman (2010) show that giving teachers feedback on their students' past performance and detailed information on how to improve students' learning in low stakes tests has no effect on tests scores, students' test scores increased when this informational treatment was paired with a monetary incentive to teachers based on the performance of their students (the treatment from Muralidharan and Sundararaman, 2011). Since there is no treatment arm with monetary incentives but no information, it is hard to disentangle whether this effect is simply due to the monetary incentives, or whether the latter are only effective when teachers are given enough information on how to influence student learning.

[55] The findings of Tran and Zeckhauser (2012) and Azmat and Iriberri (2010) are consistent with the notion that not informing students about their achievement in the ECE might keep them from applying themselves while taking the test. Both studies find that providing students with relative performance feedback enhances their performance, even if they are not rewarded for it.

[56] The evidence provided by Behrman et al. (2015) on this point is only suggestive because, as compared to the treatment in which only teachers were incentivized, the potential reward for teachers and students was larger in the treatment arm in which both were incentivized. It is therefore hard to tease out if this incremental effect is coming the existence of complementarities between teachers' and students' effort, or from the fact that the monetary incentives were larger.

our online survey to teachers. When asked whether they thought students put any effort when taking the ECE test, 37% of survey respondents who taught math or language in 8^{th} grade answered that they did not. We inquired about the reasons for why students do not put any effort while taking the ECE, and teachers replied that this was due to the fact that ECE test scores do not affect their final grade (51%), because students are unmotivated (47%), the test is too long (10%) or too difficult (8%), and students are not familiar with these types of evaluations (11%).

Since the ECE was implemented in secondary schools for the first time in 2015, schools might not have known the relative standing of their students in comparison to those from competing schools. Teachers might have been unable to infer the level of effort needed for their school to win the bonus, thus lowering the power of the incentive and ultimately discouraging them from putting in more effort. Since schools participating in the BE compete against other comparable schools within their district, they might have some prior about how their students compare to those of the competing schools, especially in BE groups with few schools. As shown in columns (5) and (10) of Table 2.9, there is a small but positive effect (0.012 and 0.023 SD in math and language) on student learning in schools in BE groups smaller than 27, the median group size. Although group size is probably an inaccurate proxy for knowledge about the probability of winning, this suggests that it might be important for schools to know how much effort they need to exert for the program to be effective.[57]

2.7.6 Teachers Did Not Have Enough Time to React

Finally, schools might not have had enough time to increase their students' learning in a meaningful way. As explained in Section 2.2, the Minister of Education mentioned the possibility of extending BE to secondary schools at the end of 2014,

[57] One of the questions in our survey shows a hypothetical ranking of 20 urban schools from the same school district, and asks teachers to identify what position would be held by a school with the same characteristics as the one they work for, and how that position would change if every teacher in their school dedicated an extra hour a day to improve the performance of their students (extra tutoring sessions, training sessions, etc.). In 47 % of cases, math and language teachers in 8^{th} grade answered that their school would still be below the 80^{th} percentile (i.e., would not win the bonus) after everyone changed their pedagogical practices.

but the programs' regulation and the Ministry of Education's corresponding diffusion campaign only came out in July 2015, four months before the November 2015 ECE. In our survey to teachers, of those who taught 8^{th} grade and knew about BE in 2015, 39% reported that they heard about the program in the first trimester, 26% in the second, and 33% in the third (and 2% could not remember when they found out). The programs implemented by Muralidharan and Sundararaman (2011), Glewwe et al. (2010) and Lavy (2009) were announced 7-8 months before students were tested. Even though these papers find positive and sizable effects in this short time frame, teachers in Peru might have had less time to react, especially those who found out about the program later in the year. Furthermore, when asked whether they thought there was enough time for students to improve their performance in the ECE in 2015 from the moment they found out about BE until the test, 67% of 8^{th} grade teachers who knew about the BE answered that there was not enough time.[58]

2.8 Conclusion

Can tying teachers' pay to the performance of their students improve their learning? We examine the impact of a collective teacher pay-for-performance program (*Bono Escuela*) implemented in 2015 in all public secondary schools in Peru, and find that it had no impact on students' math and language internal grades. Our coefficients are precisely estimated, allowing us to reject effects larger than 0.017 standard deviations, well below those previously found in the literature. Moreover, we find no evidence that the teacher incentive program had differential effects over schools or students of certain characteristics. We stipulate that the lack of increase in student learning might have been triggered by certain aspects of the evaluation linked to the bonus (students' low stakes and teachers' inexperience with it). These factors, along with schools' uncertainty about their potential ranking might have discouraged teachers from exerting higher effort.

[58] While this could be an ex-post justification, we cannot discard that lack of time is one of the reasons the program had no impact. The future analysis of students' performance in the 2016 wave of the BE, for which teachers have the entire school year to prepare, might allow us to elucidate if this is potentially one of the reasons why the program had no effect on 8^{th} graders' performance in 2015.

Finally, we argue that the program's timing might have played a role, possibly leaving teachers with insufficient time to instill significant learning gains in their students.

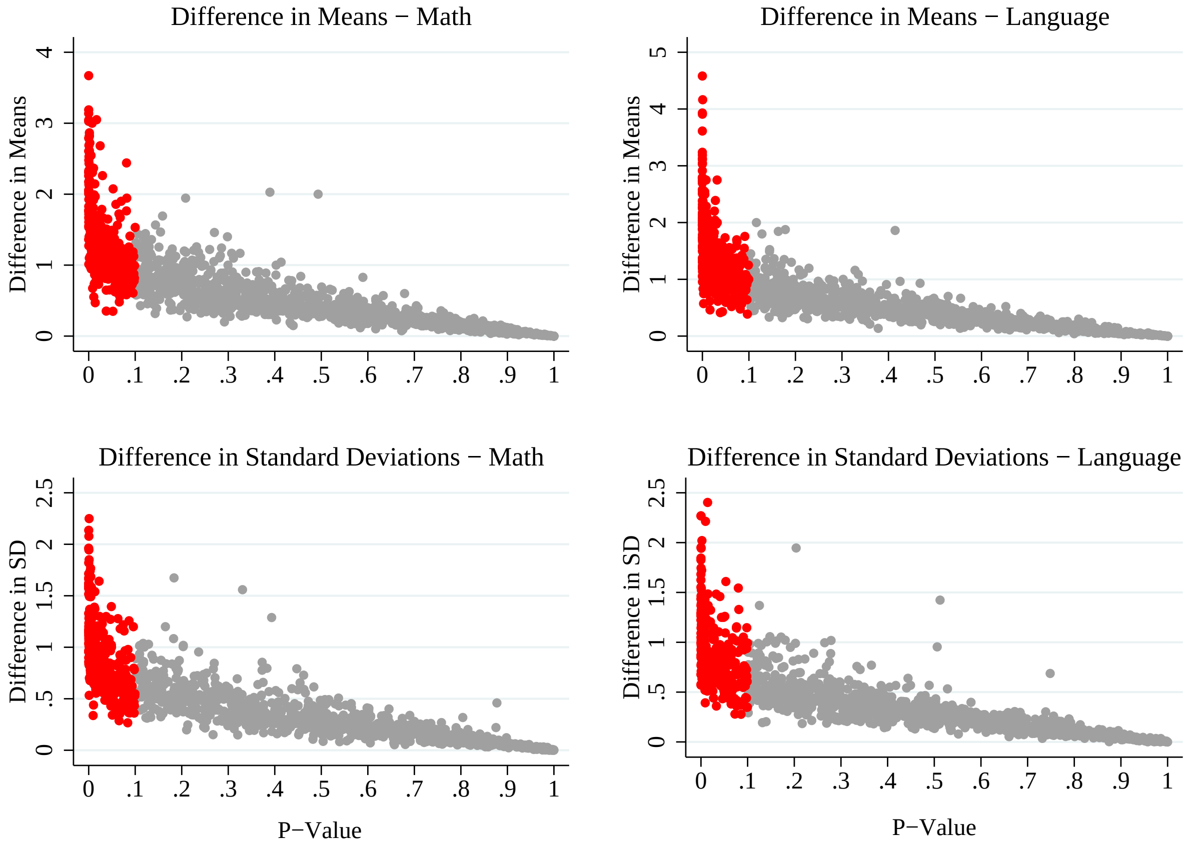
All in all, the results from our study suggest that successfully scaling up teacher pay-for-performance requires a deeper understanding about the role played by the different characteristics of these programs in their success. In particular, our findings raise the question of whether the interaction between between teachers' incentives and their information is important for these programs to work. If these complementarities exist, the efficacy of teacher incentives might depend on whether they are paired with teacher training. This paper also points to the fact that the type of exam being incentivized, and particularly the stakes that students have in it, might be important for teacher pay-for-performance programs to raise student learning. Going forward, research on teacher incentives should experimentally examine the complementarities between teachers' incentives, their knowledge, and their students' stakes in the incentivized outcome.

The fact that BE had no effect in the short-term does not imply that the program would have the same learning impacts if extended for a longer period. For one, teachers would acquire more experience with both the ECE and the BE. Consequently, some of the potential issues that could be diluting the effect of the incentive program may disappear. For instance, teachers would have more time to react to the incentive, would be more familiar with the test, and schools would have more information about their potential ranking within the group of schools they are competing against. Furthermore, teachers might only find it worthwhile to make sizable investments in improving their pedagogy if the program is continued and not only a one-off event. On the other hand, the program could have undesirable long-run impacts if teachers become more acquainted with how to teach-to-the-test, or if schools divert resources away from students not reached by the ECE. Extending the program could also result in schools devoting higher effort to improving the learning of 7^{th} grade students, in anticipation of their participation in the ECE standardized test in the following year. We plan to study these issues in our future research, once students' achievement data from 2016 be-comes available. Finally, the program could affect the quality of teachers attracted to public schools, impacting the performance of students in the entire school. Although there is some evidence on the role of financial incentives in shaping the attributes of candidates for public sector jobs (e.g., Dal Bó et al., 2013 and Deserranno, 2016),[59] this question has not been tackled in the context of teacher pay-for-performance programs yet.

[59] While Dal Bó et al. (2013) find that higher wages for advertised government jobs in Mexico attract candidates with higher capabilities and greater motivation for working in the public sector, Deserranno (2016) finds that higher financial incentives for health promoters in Uganda attract more candidates, but hamper retention and performance because people drawn to the position are less likely to have pro-social preferences.

Appendix Figures and Tables

Figure 2.A1: Variation in Internal Grades Across 8^{th} Grade Classes in 2014



Notes: Top figures depict the difference in means and the corresponding p-value when testing whether the mean math and language internal grades differ across 8^{th} grade classes from the same school in 2014, in every school with only two 8^{th} grade classes. The bottom figures depict the difference in the standard deviation and the corresponding p-value for an F-test of difference in variances in the same sample of schools.

Table 2.A1: Test for Parallel Trends Comparing Public and Private Schools

	Math	Language	Non-Incentivized Courses
Public * Post	0.116***	0.100***	0.143***
	(0.010)	(0.011)	(0.007)
Public * 2014	0.044***	0.063***	0.036***
	(0.008)	(0.010)	(0.006)
Repeated last year	-0.527***	-0.586***	-0.548***
	(0.006)	(0.006)	(0.005)
Male	0.109***	0.323***	0.271***
	(0.003)	(0.003)	(0.002)
Foreigner	0.037***	0.058***	0.061***
	(0.012)	(0.011)	(0.009)
Spanish is native tongue	0.183***	0.193***	0.126***
	(0.007)	(0.007)	(0.005)
Has a disability	-0.247***	-0.265***	-0.232***
	(0.014)	(0.015)	(0.013)
Father is alive	0.073***	0.075***	0.068***
	(0.003)	(0.004)	(0.003)
Mother is alive	0.031***	0.038***	0.039***
	(0.006)	(0.006)	(0.005)
Father lives in HH	0.037***	0.031***	0.030***
	(0.002)	(0.002)	(0.002)
Mother lives in HH	0.011***	0.016***	0.011***
	(0.002)	(0.002)	(0.002)
Observations	1514619	1514593	1514717
R^2	0.155	0.196	0.293

Notes: The sample includes all 8^{th} grade students in 2013-2015, in public and private schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. All regressions include year fixed effects, school fixed effects, and the standard controls. The dependent variables are students' internal grades in math, language and non-incentivized courses, standardized by course-year. We take the average of non-incentivized courses, which are art, science, social studies, English, civics, human relations, physical education, religion, and education for the workforce. *Public* is a dummy for whether the students attends a public school, *Post* is a dummy for the year 2015, and *2014* is a dummy for the year 2014. The coefficient for the *Public* dummy is not display since this variable is perfectly collinear with the corresponding school fixed effect. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2.A2: EFFECT OF TEACHER INCENTIVE ON STUDENTS' GRADES IN NON-INCENTIVIZED COURSES

	Arts	Science	Social Studies	English	Civics	Human Relations	Physical Education	Religion	Educ. for the Workforce
8^{th} Grade * Post	0.003	-0.001	0.017*	0.011	0.008	0.019**	0.012	0.014*	0.012
	(0.009)	(0.008)	(0.009)	(0.008)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)
8^{th} Grade	-0.037***	0.052***	-0.039***	-0.046***	-0.055***	-0.044***	-0.016***	-0.042***	-0.027***
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
Repeated last year	-0.621***	-0.570***	-0.602***	-0.573***	-0.598***	-0.588***	-0.617***	-0.611***	-0.624***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)	(0.007)	(0.008)
Male	0.336***	0.236***	0.259***	0.281***	0.339***	0.384***	0.074***	0.393***	0.256***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)
Foreigner	0.029*	0.057***	0.039**	0.163***	0.041**	0.030	0.061***	-0.015	0.003
	(0.017)	(0.018)	(0.020)	(0.019)	(0.020)	(0.018)	(0.015)	(0.018)	(0.018)
Spanish is native tongue	0.088***	0.148***	0.128***	0.147***	0.128***	0.138***	0.092***	0.103***	0.107***
	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.006)	(0.006)	(0.007)	(0.007)
Has a disability	-0.188***	-0.261***	-0.229***	-0.276***	-0.235***	-0.239***	-0.257***	-0.185***	-0.219***
	(0.016)	(0.014)	(0.016)	(0.014)	(0.015)	(0.014)	(0.015)	(0.016)	(0.016)
Father is alive	0.064***	0.063***	0.065***	0.068***	0.063***	0.066***	0.052***	0.062***	0.064***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)
Mother is alive	0.050***	0.043***	0.040***	0.050***	0.038***	0.043***	0.047***	0.039***	0.043***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)
Father lives in HH	0.024***	0.026***	0.024***	0.020***	0.025***	0.023***	0.013***	0.022***	0.023***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Mother lives in HH	0.012***	0.014***	0.013***	0.015***	0.014***	0.014***	0.012***	0.016***	0.010***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	2108795	2108792	2108780	2108791	2108793	2108777	2108794	2071576	2108778
R^2	0.196	0.127	0.138	0.143	0.154	0.168	0.235	0.183	0.186

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. The dependent variables are students' internal grades in art, science, social studies, English, civics, human relations, physical education, religion, and education for the workforce, standardized by course-year. 8^{th} *Grade* is a dummy for whether the students is in 8^{th} grade, and *Post* is a dummy for the year 2015. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2.A3: Non-Linear Heterogeneous Effects by Students' Lagged Grade

	Math	Language
Panel A: Lagged Grade Quartiles		
8^{th} Grade * Post	-0.003	0.003
	(0.009)	(0.011)
8^{th} Grade * Post * Q2	-0.006	0.009
	(0.009)	(0.010)
8^{th} Grade * Post * Q3	-0.010	0.008
	(0.010)	(0.011)
8^{th} Grade * Post * Q4	0.014	0.001
	(0.014)	(0.015)
Observations	1382813	1382756
R^2	0.393	0.387
P-value (sum of coefficients Q2)	0.346	0.281
P-value (sum of coefficients Q3)	0.198	0.349
P-value (sum of coefficients Q4)	0.380	0.792
Panel B: Lagged Grade Terciles		
8^{th} Grade * Post	-0.004	-0.004
	(0.009)	(0.010)
8^{th} Grade * Post * T2	-0.014	0.010
	(0.009)	(0.009)
8^{th} Grade * Post * T3	0.014	0.011
	(0.013)	(0.012)
Observations	1382813	1382756
R^2	0.359	0.363
P-value (sum of coefficients T2)	0.050	0.622
P-value (sum of coefficients T3)	0.444	0.588

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2014-2015, in public schools eligible for taking the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. We exclude the year 2013 for which students' previous grade is unavailable. The dependent variables are students' internal grades in math and language, standardized by course-year. Students in Panel A (B) are divided into quartiles (terciles) according to their lagged grade (i.e., their internal grade in that particular course in the previous year, standardized by school and grade). 8^{th} *Grade* is a dummy for whether the students is in 8^{th} grade, and *Post* is a dummy for the year 2015. All regressions include school and year fixed effects, as well as the standard individual controls and the three-way interaction between 8^{th} *Grade*, *Post* and the Quartile or Tercile dummies. We only report the triple interactions for exposition purposes. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2.A4: CROSS-SECTIONAL CORRELATION BETWEEN AVERAGE ECE TEST SCORES AND INTERNAL GRADES IN 2015

	Math ECE		Language ECE	
	Average (z-score)	% Very Low	Average (z-score)	% Very Low
Average internal grade (z-score)	0.116***		0.103***	
	(0.010)		(0.008)	
Failed course (% of students)		0.071***		0.091***
		(0.018)		(0.021)
Spanish as native tongue (% of students)	0.667***	-0.180***	0.836***	-0.250***
	(0.044)	(0.012)	(0.039)	(0.012)
Mother with high education (% of students)	0.930***	-0.223***	1.143***	-0.221***
	(0.078)	(0.020)	(0.065)	(0.018)
Father with high education (% of students)	0.438***	-0.120***	0.558***	-0.175***
	(0.075)	(0.020)	(0.060)	(0.018)
Repeated last year (% of students)	-0.138	0.044	0.079	-0.053
	(0.156)	(0.038)	(0.118)	(0.039)
Male (% of students)	-0.171***	0.037***	0.086*	-0.029**
	(0.054)	(0.014)	(0.046)	(0.013)
Teacher-pupil ratio	0.010***	-0.002***	0.010***	-0.002***
	(0.001)	(0.000)	(0.001)	(0.000)
Long school-day	0.309***	-0.074***	0.187***	-0.052***
	(0.024)	(0.006)	(0.018)	(0.005)
Rural	-0.050**	0.019***	-0.135***	0.049***
	(0.023)	(0.006)	(0.018)	(0.006)
Observations	8010	8010	8010	8010
R^2	0.501	0.491	0.684	0.617

Notes: The sample includes all public secondary schools taking the ECE in 2015 and registered in the Ministry of Education's SIAGIE administrative system, and the unit of observation is a school in the year 2015. The dependent variable in columns 1 and 3 are 8^{th} graders' average ECE grades in math and language, expressed as a z-score. The dependent variable in columns 2 and 4 is the % of students with very low achievement in the 2015 ECE. *Average internal grade (z-score)* is the school's average internal grade for 8^{th} grade students in 2015 , standardized across schools, in math (column 1) and language (column 3). *Failed course* measures the % of 8^{th} graders that failed math (column 2) and language (column 4) in 2015 according to their internal grades. *Mother with high education* and *Father with high education* indicate the % of 8^{th} graders in that school whose mother/father had more than a primary school degree in 2015. *Teacher-pupil-ratio* is the average number of 8^{th} grade students per class in 2015, and *Long school-day* is a dummy for whether the school had a longer instruction day in 2015. All regressions include school district fixed effects. Robust standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2.A5: Cross-Sectional Correlation Between School Covariates and Average Learning Outcomes in 2015

	Math		Language	
	ECE	Internal	ECE	Internal
Spanish as native tongue (% of students)	0.672***	0.044	0.857***	0.206***
	(0.045)	(0.056)	(0.039)	(0.054)
Mother with high education (% of students)	0.959***	0.252***	1.180***	0.363***
	(0.079)	(0.089)	(0.066)	(0.092)
Father with high education (% of students)	0.474***	0.314***	0.591***	0.327***
	(0.076)	(0.084)	(0.061)	(0.083)
Repeated last year (% of students)	-0.438***	-2.578***	-0.206*	-2.772***
	(0.157)	(0.168)	(0.117)	(0.170)
Male (% of students)	-0.132**	0.334***	0.125***	0.385***
	(0.055)	(0.065)	(0.046)	(0.065)
Teacher-pupil ratio	0.008***	-0.023***	0.008***	-0.020***
	(0.001)	(0.002)	(0.001)	(0.002)
Long school-day	0.338***	0.242***	0.214***	0.269***
	(0.024)	(0.030)	(0.018)	(0.030)
Rural	-0.048**	0.023	-0.135***	-0.001
	(0.023)	(0.028)	(0.018)	(0.028)
Observations	8010	8010	8010	8010
R^2	0.491	0.243	0.676	0.239

Notes: The sample includes all public secondary schools taking the ECE in 2015 and registered in the Ministry of Education's SIAGIE administrative system, and the unit of observation is a school in the year 2015. The dependent variable in columns 1 and 3 are 8^{th} graders' average ECE grades in math and language, whereas the dependent variable in columns 2 and 4 are the school's average internal grade for 8^{th} grade students in 2015. Average ECE and internal grades are standardized across schools (i.e., expressed as a z-score).*Mother with high education* and *Father with high education* indicate the % of 8^{th} graders in that school whose mother/father had more than a primary school degree in 2015. *Teacher-pupil-ratio* is the average number of 8^{th} grade students per class in 2015, and *Long school-day* is a dummy for whether the school had a longer instruction day in 2015. All regressions include school district fixed effects. Robust standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2.A6: Within-School Correlation Between Covariates and Learning Outcomes in 2015

	Math		Language	
	Grade (z-score)	Low Achievement	Grade (z-score)	Low Achievement
Panel A: ECE Grades				
Socioeconomic status index	0.135***	-0.044***	0.190***	-0.049***
	(0.003)	(0.001)	(0.003)	(0.001)
Male	0.220***	-0.071***	0.020***	-0.010***
	(0.004)	(0.002)	(0.004)	(0.002)
Observations	354429	354547	354529	354547
R^2	0.020	0.216	0.015	0.254
Panel B: Internal Grades				
Spanish as native tongue	0.204***	-0.016***	0.218***	-0.011***
	(0.012)	(0.004)	(0.012)	(0.003)
Mother with high education	0.158***	-0.021***	0.166***	-0.014***
	(0.005)	(0.002)	(0.005)	(0.001)
Father with high education	0.129***	-0.017***	0.136***	-0.014***
	(0.005)	(0.002)	(0.005)	(0.001)
Male	0.138***	-0.035***	0.353***	-0.051***
	(0.005)	(0.002)	(0.005)	(0.001)
Observations	324696	325320	324689	325320
R^2	0.019	0.103	0.044	0.099

Notes: Panel A contains the anonymized sample of 8^{th} graders taking the ECE standardized test in 2015, and the sample from Panel B includes all 8^{th} graders in 2015 from public secondary schools taking the ECE in 2015 and registered in the Ministry of Education's SIAGIE administrative system. The dependent variable in columns 1 and 3 of Panel A (Panel B) is the students' ECE (internal) grade in math and language, standardized by subject and school. The dependent variable in columns 2 and 4 of Panel A (Panel B) is a dummy for whether the student scored in the lowest category in the ECE (failed according to his internal grades). *Socioeconomic status* is an index ranging between -3.5 and 9.5, which is increasing in measures of socioeconomic status such as parents' education, and household assets and characteristics. All regressions include school fixed effects. Standard errors clustered by school in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Chapter 3

The Effect of Increasing the Legal Working Age on Women's Fertility and Infant Health

Joint with Sergi Jiménez-Martín and Judit Vall-Castelló

3.1 Introduction

Decreasing fertility rates is one factor contributing to the ageing of the population, a major concern in many industrialized countries due to the increased pressure on the sustainability of their social security systems. Many countries in 2005-2010 have a total fertility rate (TFR) below replacement, 2.1 births per woman. For instance, the TFR is 1.5 births per woman in Europe and 1.4 births per woman in Japan. Lately many governments enacted policies to raise fertility in order to slow population aging.[1] Not only total fertility rates are important but also health outcomes at birth are crucial to determine future outcomes of affected children. It has been well established in the literature that infant health outcomes have long-term consequences in terms of cognitive development (Figlio et al., 2014), adult health (Fletcher et al., 2010) and productivity (Smith, 2009).

[1] *Source*: Lee et al. (2014). For instance, France has increased its fertility rates from 1.74 to 2.08 through some pro-natalist initiatives, such as tax deductions for dependents and paid maternity leave financed through the national health insurance system.

Some researchers have pointed to the role of education in explaining the reduction in fertility rates and improvements in infant health outcomes. An extensive literature has examine the causality between education and fertility (León, 2006; Black et al., 2008; Fort, 2007; Monstad et al., 2008; Fort et al., 2011; Silles, 2011; Cygan-Rehm and Maeder, 2013; Geruso et al., 2014), marriage (Kırdar, 2009), or infant health (Behrman and Rosenzweig, 2002; Currie and Moretti, 2003; McCrary and Royer, 2011) using changes in compulsory schooling laws as a source of exogenous variation on individual schooling choices. Another part of the literature investigates the impact of marriage law changes on poverty (Dahl, 2010) fertility, and schooling outcomes (Bharadwaj, 2015; Buckles et al., 2011).

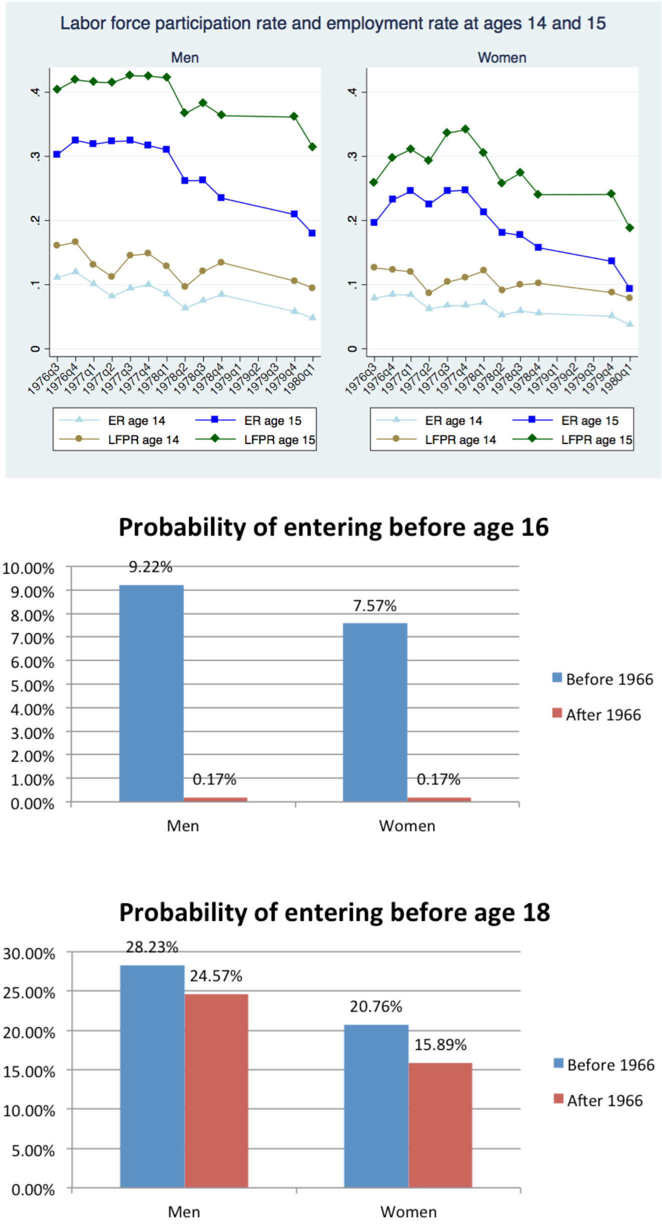
In contrast, the effect of child labor laws affecting the minimum legal age to work on long-term outcomes has been fairly overlooked. Goldin and Katz (2011), Lleras-Muney (2002) and Edmonds and Shrestha (2012) examine the effect of child labor laws, jointly with other compulsory schooling laws, on educational attainment. Contrary to our setting, the minimum age to work was normally set at a lower age than the maximum age of compulsory schooling. Goldin and Katz (2011) find that compulsory schooling and child labor laws in the US from 1910 to 1939 only modestly increased high school enrollments. On the other hand, Lleras-Muney (2002) finds that these laws increased educational attainment of those individuals in the lower percentile of the distribution. Edmonds and Shrestha (2012), that analyze the effects of minimum age of employment in 59 low income countries, find that these laws were barely enforced.

As far as we are aware of, this is the first paper that investigates the effect of a child labor regulation that increased the legal age to work on fertility and infant health outcomes. For that, we exploit a reform that introduced an exogenous variation in the Spanish legal age to work. This strategy is in stark contrast to the previous literature that has examined the role of education over these family behavior outcomes using compulsory schooling laws. We exploit the interaction between the compulsory schooling age and the minimum legal age to work to identify the incentives of different types of individuals. We argue that both age thresholds affect the decision to remain in the educational system so that it is important to consider both factors at the same time. Finally, we claim that increasing the minimum legal working age could be a more efficient and cost-reducing way of increasing educational attainment as opposed to an increase in the number of years of compulsory schooling. Thus, our paper provides evidence of a potential policy alternative for a number of developing countries.[2]

We take advantage of a quasi-natural experiment. In 1980, a new child labor regulation was enacted, the Workers Statute (Law 8/1980), which changed the minimum legal age to work in Spain from 14 to 16 years old. This reform took place a few years after the democratization of Spain following the end of the dictatorial regime. Thus, at that time, Spain was still a developing country with a large percentage of its population achieving low levels of education while participating in the labor market from an early age. The top panel of Figure 3.1 displays the labor force participation and employment rates for children of 14 and 15 years old, five years before the implementation of the reform. In 1976, around 40 percent of males and 30 percent of females were already participating in the labor market at age 15. This percentage remained high until 1980. Participation rates for children age 14 were lower. Around 15 percent of males and 10 percent of females were working at age 14 during the early 1970s. These percentages dropped to 10 and 5 percent, respectively, in the first quarter of 1980 (the last period for which we have estimates of labor force participation for children under 16 years old from the Labor Force Survey). The bottom panels of Figure 3.1 show the age of the first Social Security contribution for pre- and post-reform cohorts. Prior to the reform, for the cohorts born between 1961 and 1965, 9.22 percent of boys and 7.57 percent of girls started working in the formal labor market before age 16. After the reform, these numbers drop dramatically, with almost no one contributing before the age of 16.

[2] The reform in question was implemented when Spain was still a developing country. Thus, our results are particularly relevant from a policy perspective for countries whose level of education and child labor market participation are similar to the levels Spain was experiencing during the early 1980s (reported in Figure 3.1).

Figure 3.1: Labor Force Attachment and the Age of Labor Market Entry



Notes: ”Before 1966” refers to the cohorts born in 1961-1965. ”After 1966” refers to the cohorts born in 1967-1971. Number of observations: Men: 123,050; Women: 108,483. *Source*: Spanish Labor Force Survey and Muestra de Condiciones de Vida Laboral (MCVL).

One could think that a substantial part of this employment was in the informal market and, thus, not captured in the Spanish Labor Force Survey or the Social Security contributions. The Spanish Household Budget Survey of 1980-81 (Alonso-Colmenares et al., 1999) reveals that, after the reform, only 2.1 percent of the boys and 1.2 percent of the 14-years-old girls were participating, formally or informally, in the labor market. Likewise, 9.63 (5.1) percent of the 15-years-old boys (girls) were participating in the labor market in 1980. Thus, one third of the employment of children under 16 years old was in the formal market, and the reform not only eliminated child formal work, but also reduced the informal part of child employment (under 16 years old).

We use a differences-in-differences strategy to identify the reform's within-cohort effects. In our setup, treated individuals (born between January and May) and their control counterparts (born between July and December)[3] only differ in their month of birth. Consequently, our identification strategy improves upon the before-after analysis commonly used in the literature (which would identify the reform's between-cohort effects). Any concurrent social or political event that might have affected the post-reform cohorts would have had a similar impact on both our treatment and control groups. Unlike most of the extant literature, we use registered data of all births and marriages in Spain, which allows us to observe the universe of all birth and marriages that took place during more than 30 years.[4] This type of data has some advantages over census data, which only identifies a woman's children as those living in the same household at the time of the interview. Divorce, death of the mother, or the emancipation of older children can have an impact on this number. If the level of education affects the probability that some of these situations occur, then census data could bias the results.

We find that the reform significantly increased the women's probability of remaining childless and reduced their completed fertility, with 1,786 fewer

[3] Note that we deliberately exclude the month of June, as it coincides with the end of the academic year.

[4] See the data appendix for a description of all datasets and register data used in this paper.

women having children and 3,307 fewer children being born in the 10 cohorts following the reform. These effects operate through a postponement of first births until an age when the catching-up effect cannot take place due to decreased fertility with age. We also show that the marriage market is another channel through which the reform affects fertility by delaying the age at which women first marry and reducing the likelihood that a woman ever marries.

Importantly, we show that the postponement in fertility is detrimental for the health of their offspring at the moment of delivery. The reform caused 2,789 more children to be born at less than 37 weeks of gestation and 4,352 were born with low birth weight. We propose three different channels through which the reform could be negatively impacting infant health. The first is the postponement of the age at which women have their first child, which increases the probability of having this first child after the age of 35. Because the risk to infant health during pregnancy increases after that age, postponing motherhood translates into negative effects on infant health outcomes. We also show that the reform changed the maternal marital status, increasing the number of children whose mother were not married or that had no registered father. In addition, we show that better employment prospects of more educated women enhances unhealthier behaviors (smoking and drinking), further contributing to the negative effects that we report on infant health outcomes.

This last result may seem surprising given prior research showing a negative relation between years of schooling and smoking prevalence among women in developed countries (for instance, Currie and Moretti (2003)). This finding can be explained by differences in labor market integration and educational attainment between men and women in the pre-reform cohorts. For example, Bilal et al. (2015) find a substantial difference in smoking prevalence by gender in cohorts born in Spain between 1940 and 1960 (pre-reform cohorts), with highly educated women having the highest smoking prevalence rates and women with fewer years of education exhibiting lower rates of smoking. This inverse gradient for Spanish women is gradually reversed until the cohorts of women born after 1980, when the country's gradient begins to mirror that of developed countries, with less educated women showing the highest smoking prevalence rates.

More importantly, this positive gradient between education and smoking for women is not unique to this setting. Previous research has established a higher smoking prevalence among high educated women than among low educated women in Eastern Europe and Eastern Mediterranean countries (Bosdriesz et al., 2014). Also, Pampel (2003) showed that high-income countries at early stages of the smoking epidemic (like the southern European countries a few decades ago) had higher rates of female smokers among the young and highly educated. This is due, primarily, to the weakening of the social and cultural constraints that prevented many women from smoking in the past (Mackay and Amos, 2003).

The remainder of the paper is organized as follows. Section 3.2 presents the institutional context as well as the identification strategy. In Section 3.3 we present the effects of the reform on fertility rates and infant health outcomes. In Section 3.4, we perform a number of robustness checks while Section 3.5 concludes with a discussion of the main results and their policy implications. At the end of the paper, a data appendix can be found, with a detailed explanation of all the databases used in the paper, as well as a broader before-after analysis of the reform.

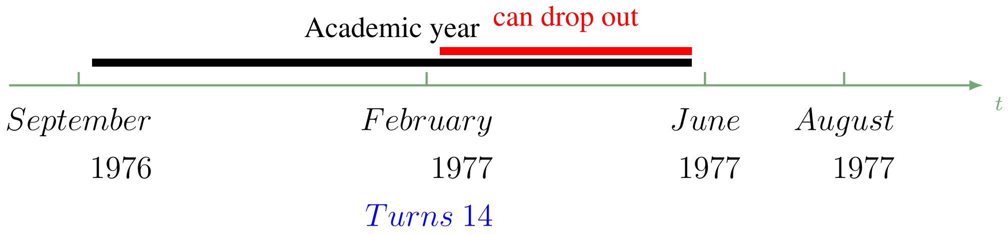
3.2 Institutional Context

Our identification strategy builds on an exogenous variation in the incentive to stay out of the labor market induced by a legislative change in the legal age to work in Spain. Law 8/1980 "Estatuto de los Trabajadores" (ET) was introduced in March of 1980 as a child labor law that increased the minimum legal working age from 14 to 16 years old. Only individuals born after 1966, who were 14 at the time the reform was passed, were subject to the reform. Therefore, we compare individuals who turned 14 just after the reform to those who turned 14 just before the reform.

Additionally, not all individuals from the same cohort were affected by the reform in the same way. Before the reform, students born during the first months of the year reached the minimum legal working age of 14 before finishing their last year of primary education.[5] Therefore, they had an incentive to leave school before completing primary education. On the other hand, students born during the last months of the year had reasons to finish primary education, as they were not old enough (had not turned 14 years old) to legally work before finishing primary education. Consequently, we expect that before the reform was passed, those born at the beginning of the year would have a lower probability of finishing primary schooling than individuals born at the end of the year.

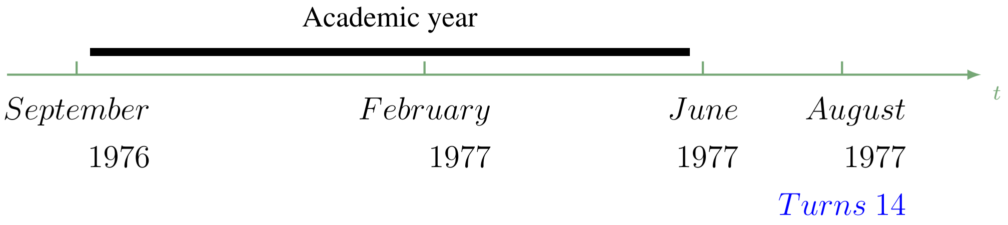
After the reform, this difference in incentives disappears. The reform increased the legal working age to 16 years old, but the compulsory schooling age remained at 14. Thus, after the reform, all individuals in the same cohort had similar incentives to complete the last year of primary education as they were not able to work until age 16. The following chart illustrates the timing of the reform by showing two individuals in the same 1963 cohort (pre-reform), during their last year of primary school:

1. An individual that was born on February of 1963:



[5] Note that in the Spanish educational system, all children from the same cohort start school the same year. Consequently, children born at the beginning of the year start school at an older age (in months) than those born at the end of the year.

2. For an individual that was born on August of 1963:



Before the reform, the two individuals' incentives to stay in the educational system during the last year of primary education differed depending on whether they were born during the first part of the year (from January to May) or the last part of the year (from July to December).

We exploit the exogenous change in the incentives introduced by the ET reform to identify the causal effect of a child labor regulation on fertility and infant health outcomes. We focus on the variation among individuals from the same cohort but born at different times of the year, before and after the reform. Thus, we are not making a before-after comparison, as the prior literature has done when analyzing the effects of changes in the educational laws. We are aware that the impact of the ET could potentially be greater than what we estimate using the within-cohort comparison and, in the appendix, we provide estimates of this before-after effect. However, in 1980, the year that the reform was introduced, Spain was experiencing a period of significant social change. The democratization process in Spain took place in 1979 and a number of reforms passed quickly thereafter. For instance, divorce was legalized in 1981 and abortion in 1985. Consequently, the cohorts of women that turned 14 years old before and after the reform are exposed to different environments. Even if we observe a significant change in fertility and marriage after the 1980 reform, this change could be due to the influence of other reforms that were taking place concurrently. Hence, our strategy is much more conservative as we are only exploiting the within-cohort variation but, in this setting, the identification strategy is much more reliable than a before-after modeling approach.

Formally, we consider the following econometric model:

$$Outcome_i = \alpha + \beta_1\, Treated_i + \beta_2\, Treated_i * Post\ Reform_i$$
$$+\beta_3\, BirthYear\ FE + \beta_4\, CalendarYear\ FE + \beta_5\, Region\ FE + \epsilon_i$$

where $Treated_i$ is a dummy variable that equals one if individual i was born between January and May and zero if she was born between July and December. $Post\ Reform_i$ is also a dummy variable that takes a value of one if individual i turned 14 after the reform and zero otherwise. Then, we define pre-reform cohorts as those born in 1961 to 1965, and post-reform cohorts as those born between 1967 and 1971. We also include region, birthyear and calendar year fixed effects and cluster the standard errors (in parenthesis) at the cohort level. Given the small number of clusters, we also perform a wild bootstrap with 1,000 repetitions and we report the p-value in brackets. The effect of the reform can be identified by the coefficient of the interaction between the post-reform and the treatment dummy variables, β_2. We will use this same econometric specification in the rest of the paper. All results are robust to the exclusion of region fixed effects and the inclusion of the interaction between the cohort and region fixed effects.

We borrow the first stage of the reform from Jiménez-Martín et al. (2015). They show that the reform was effective in providing incentives for treated individuals to, not only finish primary education, but also to continue in the educational system afterwards. In particular, they find that the increase in the minimum statutory working age decreased the number of early school leavers (individuals not finishing primary education) in 1.61 percentage points (7.16%) for men and 0.98 percentage points (7.7%) for women. At the same time, the reform decreased the number of treated individuals not attaining post-compulsory education (dropouts) by 1.82 percentage points (3.7%) for men and 1 percentage points (2.28%) for women. These results certify that the reform was effective in increasing educational attainment of affected individuals and, thus, in restricting child labor. For the rest of the paper we will analyze the effects of the reform on fertility and marriage outcomes as well as on infant health at delivery.

3.3 Effect of the Reform on Family Behavior Outcomes

3.3.1 Effect of the Reform on Fertility

We first study the impact of the reform on several fertility outcomes. To test whether women affected by the reform postpone motherhood, we first examine the impact of the reform on the age at which women have their first child. Secondly, we assess whether the reform affected the probability of women remaining childless or the number of children they have. In other words, we want to determine whether there is a catching-up effect after the delay of motherhood.

For these estimations, we use register data on all birth certificates from 1975 to 2014, available from the Spanish National Statistics Institute (see the data appendix for a detailed explanation of the birth register). Our pre-reform cohorts comprise women born between 1961 and 1965, and the post-reform cohorts include those born between 1967 and 1971. In addition, we restrict the sample to all births from women born in Spain and those births that took place when the mother was between the ages of 14 and 43. This age restriction allows us to include the same ages for all the cohorts considered, as women of the first cohort (1961) were 14 in the first year of the register and women of the last cohort (1971) were 43 in the last year of the register. We define the probability of ending childless as the ratio between the total number of first births and the total number of women born in a certain cohort and treatment status.[6] Similarly, we define completed fertility as the ratio between the total number of births and the total number of women born in a certain cohort and treatment status. Thus, to examine the effect of the reform on the probability of ending childless and completed fertility, the data has been collapsed at the cohort level, with cells defined at the level of treatment, cohort, year of birth and region.

The estimates in Table 3.1 show the effect of the reform on the age at which

[6] The results are multiplied by 1,000.

women had their first child. Before the reform, women born at the beginning of the year had their first child almost a month (28 days) earlier than women born at the end of the year. This gap in age disappears (is reduced by 21 days) after the reform is introduced.[7] Similarly, in graph a) of Figure 3.2 we can observe that the difference in the age at which women have their first child between women born at the beginning and at the end of the year is significantly negative for all the pre-reform cohorts, while this difference is not longer significant for cohorts affected by the reform.

Table 3.1 also shows that the postponement effect is followed by an increase in the probability of remaining childless as well as a decrease in completed fertility.[8] Thus, 221 more women born in 1967[9] decided to remain childless, a decrease of 0.19 percent in the number of women that have children after the reform was implemented.[10] Given that we are exploiting, on average, three additional months of education,[11] the effect of the reform over the probability of not having any children[12] is lower than León (2006) in the US (who find that an increase in an additional year of schooling raises the probability of not having any children by almost 2 percentage points), and Cygan-Rehm and Maeder (2013) in Germany (that find that an increase in an additional year of schooling rasies the probability of not having any children by 5.1 percentage points).[13]

[7] Results are robust in sign and significance to the substitution of cohort time dummies by linear, quadratic and quartic pre- and post-reform trends, the exclusion of region fixed effects, and the incorporation of the interactions of cohort and region fixed effects.

[8] Be aware that we are only considering births that took place between the ages of 14 and 43. Thus, we cannot completely rule out the catching-up effect, as this effect could be taking place after the age of 43. However, only 0.66% of women of these cohorts have their first child with more than 43 years old.

[9] or 2,198 born between 1967 and 1976

[10] Note that 115938 women born at the beginning of the year 1965 decided to have children.

[11] The introduction of the reform increased the incentives to continue five additional months of primary school for children born during January, four months for those born in February, three months for those born in March, two months for those born in April and two month for children born in May.

[12] We find that an increase of three months of schooling increases the probability of not having a child by 0.16 percentage points, which would be equivalent to an increase of 0.64 percentage points for an additional year of education.

[13] The rest of the papers in the literature do only find a postponement effect of fertility away from the teenage years (Black et al., 2008; Fort, 2007; Silles, 2011; Geruso et al., 2014).

Table 3.1: EFFECT OF THE REFORM ON FERTILITY OUTCOMES

	Age first birth (1)	Perc. women in each cohort become a mother (2)	Number of children per women in each cohort (3)
Treated	-0.080**	1.319**	3.034***
	(0.009)	(0.203)	(0.390)
	[0.023]	[0.037]	[0.004]
Treated* Post Reform	0.058***	-1.627***	-3.079***
	(0.011)	(0.123)	(0.187)
	[0.002]	[0.003]	[0.003]
Observations	2,493,107	9,982	10,026
R^2	0.067	0.345	0.373
BirthYear FE	YES	YES	YES
Calendar FE	NO	YES	YES
Region FE	YES	YES	YES
Mean pre-reform	26.83	29.03	52.55
Std. dev. pre-reform	5.507	23.40	38.23

Notes: The dependent variables are (1) the age of the woman when she had their first child, (2) the percentage of (treated and control) women that had at least one child (multiplied by 1,000), and (3) the total number of children divided by the total number of women born in each cohort (multiplied by 1,000). Regressions include cohort time and region fixed effects, and (2-3) calendar year dummies. Note that we cannot include calendar year dummies when the dependant variable is age as it takes out all the variation. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1971.

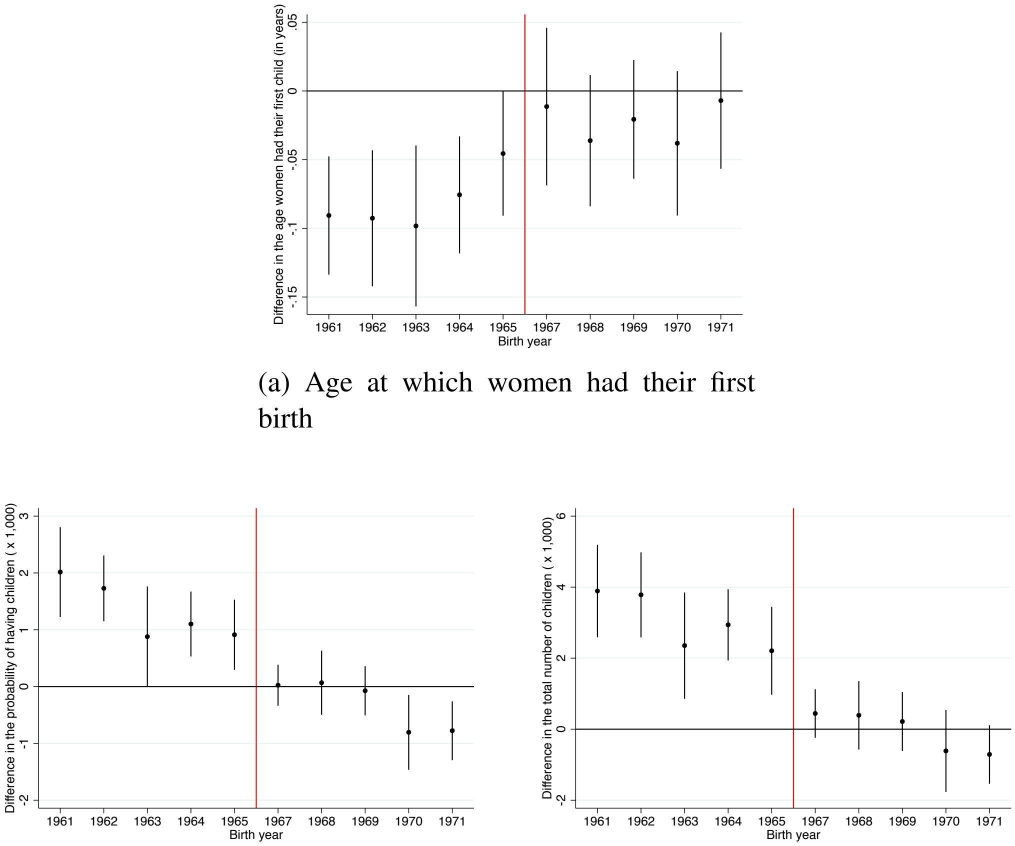
Moreover, before the reform, 3.17 more children were born per 1,000 women born at the beginning of the year with respect to women born at the end of the year. This gap is eliminated after the introduction of the reform. This means that 419 fewer children were born to the first post-reform cohort of women (born in1967)[14] corresponding to a 0.19 percent decrease in the total number of children born, given that women born at the beginning of 1965 had in total 209,954 children. This effect on completed fertility is lower to León (2006) and

[14] or 4.160 children in the subsequent 10 generations affected by the reform.

Cygan-Rehm and Maeder (2013), which find that an additional year of education reduces fertility by 0.25 and 0.1 children respectively, while the reduction on fertility as a consequence of the reform in Spain is of 0.012 children for additional year of schooling. On the other hand, the estimated effect on completed fertility is bigger than the zero effect found before by Black et al. (2008), Fort (2007), Monstad et al. (2008), Silles (2011), or Geruso et al. (2014). In graphs b) and c) of Figure 3.2 we can observe that the difference in the probability of having children or the total number of children between women born at the beginning and at the end of the year is significantly positive for all the pre-reform cohorts, while this difference is not longer significant or even negative for cohorts affected by the reform.

Fertility and completed fertility are calculated as the ratio between the first births or total number of births and the total number of women born in a certain cohort and treatment status. One potential concern is that the reform impacts mortality and migration differently for treated and control women born after the reform. If, as a consequence of the reform, the mortality rate or migration of treated women born after the reform decreased, we could incorrectly estimate a decrease in the fertility and completed fertility rates by affecting the denominator (treated women of childbearing age living in Spain) rather than the numerator (births). To address this potential concern, we test if the ratio between the treated and control women observed in census and born with in each cohort varies with the reform. In Figure 3.A1 we can observe that the proportion of treated and control women within each cohort that are represented in the Census of 2001 is quite constant around 0.048 for all the cohorts of women. Thus, we believe that differential migration and mortality rates for treated and control women is not driving the fertility results.

Figure 3.2: Differences in Fertility Outcomes of Women in the Same Cohort Born at the Beginning and at End of the Year



(a) Age at which women had their first birth

(b) Proportion woman that have children (x 1000)

(c) Total number of children per women (x 1000)

Notes: Estimated difference in (a) the age (in years) woman have their first child, (b) probability they have children, and (c) total number of children between women born at the beginning (treated) and at the end of the year (control) of each cohort. Regressions include region and calendar time dummies. The confidence intervals are calculated using standard errors clustered at cohort and region levels. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1971.

As a robustness check, we examine the probability of remaining childless and the completed fertility rate using data from the 2011 census, which includes a representative sample of 5 percent of the population and provides information about the number of children that women had up until 2011 (see data appendix for more information on the census of 2011).[15] Table 3.2 shows that the reform's effect on the probability of having at least one child and the completed fertility rate goes in the same direction as the results found using birth registries. However, the results are not significant. We believe three main factors can explain the lack of significance in the coefficients estimated with the 2011 census. First, we are only observing 5 percent of the population, so the results could be estimated with more noise, and, thus, the standard errors are higher. Second, as the census does not include information on the year in which women had their children, we cannot control for calendar year effects. Third, we only observe those children that are still in the household of their parents at the moment of the interview. It is plausible that older children from the less educated households are no longer living with their parents. This makes the census data a selected sample that biases the results. Thus, we strongly believe that the census data represent a worse database to analyze the effect of the reform on fertility outcomes.

Therefore, we conclude that the reform had two main effects on fertility. First, it made some women postpone the entrance into motherhood and this delay was not compensated for later in life.[16] Second, it increased the number of women that remained childless.

[15] Note that we are considering the same cohorts of women (1961–1971) and are defining treatment in the same way. However, in 2011, the last cohort we are considering (1971) had only reached the age of 40, so we do not observe the late births of the younger cohorts. Also, we can no use the age constraint, as we do not have information on the age of the mother when they had their children. So, we observe a higher number of births from the older (and probably less educated) women.

[16] We also evaluate this result using the 2006 "Encuesta de Fecundidad" (i.e. *Fertility Survey*), available from *Centro de Investigaciones Sociológicas* (CIS). This questionnaire was given to 10,000 women who were over the age of 15 in 2006. Here, we also have information on the total number of children that women from the cohorts of 1964 to 1968 had in 2006. The number of observations, however, is very small (around 600 women). Thus, although results go in the same direction, they are not significant.

Table 3.2: EFFECT OF THE REFORM ON FERTILITY OUTCOMES USING CENSUS DATA

	Prob. of having a child (1)	Total number of children (2)	Prob. of having 3 or more children (3)
Treated	0.003*** (0.001) [0.003]	0.017** (0.003) [0.053]	0.008** (0.001) [0.053]
Treated*Post Reform	-0.000 (0.002) [0.884]	-0.005 (0.006) [0.497]	-0.004 (0.002) [0.128]
Observations	269,392	269,392	269,392
R^2	0.009	0.025	0.016
BirthYear FE	YES	YES	YES
Calendar FE	NO	NO	NO
Region FE	YES	YES	YES
Mean pre-reform	0.819	1.598	0.130
Std. dev. pre-reform	0.385	1.019	0.336

Notes: The dependent variables are (1) the probability that a woman has at least one children, (2) total number of children per women and (3) the probability that a woman has at least 3 children. Regressions include cohort and region dummies. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Census 2011, all women from cohorts 1961-1971.

Mechanisms

This section explores the potential channels that may be preventing the catching-up effect from taking place. The main hypothesis is that the reform delays entrance into motherhood until an age after which the catching-up effect can no longer take place. To check the validity of this hypothesis, we estimate some age-specific probabilities of having the first birth. More precisely, we use the same econometric model as before but the outcome now is the probability of having the first birth at a certain age bracket. We choose the age brackets to all have the same number of years (5 years).

Table 3.3 reveals no significant effect of the reform on the probability of having the first child during the teenage years (between 14 and 18 years old). In contrast to the findings of prior studies, this evidence indicates that the reform did not induce a postponement of first births away from the teenage years. Thus, we can rule out the "incarceration effect"[17] as the main channel through which the child labor reform affected fertility. On the other hand, the reform did affect the probability of having the first child after age 29. Our results show that, after the reform, affected women had a lower probability of having their first child between the ages of 29 and 33, as well as a higher probability of having their first child after the age of 34. Therefore, we conclude that the reform decreased the probability of pregnancy during the early thirties and increased the probability of having late first births.[18]

Even if the postponement of 21 days on average seems like a small effect, the increase in the incidence of first births after the age of 34 is not. The medical literature has shown that after age 35 a woman's fertility decreases. Thus, catching-up may no longer be possible for some women, causing the observed decrease in completed fertility rates.

The reduction and postponement of fertility may be the result of a similar postponement and reduction of marriage. Note that in the 1980s in Spain, the majority of mothers had their children during their marriage. In fact, 88.86% of mothers were married when they had their first child. Thus, as an additional potential factor that may help explain the effects of the child labor reform on fertility, we analyze whether the reform had any impact on marriage outcomes. First, we study whether the reform induced women to postpone the age at which they marry for the first time. Next, we examine if this postponement reduces the number of first and total marriages over time.

[17] We define "incarceration effect" as a delay in fertility for the additional amount of time that women stay in school.

[18] Results are robust in sign and significance to the substitution of cohort time dummies by linear, quadratic and quartic pre- and post-reform trends, the exclusion of region fixed effects, and the incorporation of the interactions of cohort and region fixed effects.

Table 3.3: Effect of the Reform on the Probability of Having the First Birth at a Certain Age Bracket

	Probability of having the first birth					
	Between 14 and 18	Between 19 and 23	Between 24 and 28	Between 29 and 33	Between 34 and 38	Between 39 and 43
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	0.001*	0.007**	-0.003*	-0.003**	-0.001*	-0.000*
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
	[0.061]	[0.024]	[0.076]	[0.018]	[0.087]	[0.057]
Treated* Post Reform	-0.001	-0.002	0.002	-0.004***	0.003***	0.001***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.000)
	[0.223]	[0.232]	[0.279]	[0.009]	[0.001]	[0.001]
Observations	2,493,107	2,493,107	2,493,107	2,493,107	2,493,107	2,493,107
R^2	0.010	0.031	0.011	0.023	0.019	0.005
BirthYear FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Mean pre-reform	0.0690	0.258	0.339	0.226	0.0874	0.0211
Std. dev. pre-reform	0.253	0.437	0.473	0.418	0.282	0.144

Notes: The dependent variables are the probability of having a first child between the ages of (1) 14 and 18, (2) 19 and 23, (3) 24 and 28, (4) 29 and 33, (5) 34 and 38, and (6) 39 and 43. Regressions include cohort and region dummies. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1971.

For this analysis, we use register data on all marriage certificates from 1976 to 2012 (see the data appendix for a detailed explanation of the marriage register). As before, we consider the 1961–1965 cohorts to be pre-reform and the 1967–1971 to be post-reform. We restrict the sample to all marriages that took place when the woman was between the ages of 15 and 41. The definition of treatment and control is the same as before. Finally, we drop same-sex marriages due to their late acceptance in the definition of marriage.

For the analysis of the impact of the reform on the number of total marriages, we collapse the data at the cohort and calendar-year level for the treatment and control groups and divide them by the total number of women born to a certain cohort and treatment status. Similarly, to calculate the probability of having never married we divide the total number of first marriages by the total number of women born to a certain cohort and treatment status.

Table 3.4 shows the effects of the reform on the age at the time of marrying. Before the reform, women born at the beginning of the year married, on average, almost half a month earlier than women born at the end of the year. This difference in age between women of the same cohort is almost entirely eliminated after the reform is introduced.[19] Note that this postponement in the age of first marriage is almost identical to the postponement in the age of having a first birth.

Using data collapsed at the cohort level, with cells defined at the level of treatment, cohort, year of marriage and region, Table 3.4 also reveals that the postponement in marriage is accompanied by an increase in the probability of remaining single as well as a decrease in the total number of marriages per woman. After the reform, more than one in every 1,000 women born at the beginning of the year never married. Moreover, we observe a similar reduction in the total number of marriages per woman.[20]

[19] Results are robust in sign and significance to the substitution of cohort time dummies by linear, quadratic and quartic pre- and post-reform trends.

[20] However, these two results should be taken with caution, as we are only considering marriages that took place between the ages of 15 to 41. Thus, we cannot conclude that there is no catching-up effect, if this effect takes place after age 41.

Table 3.4: EFFECT OF THE REFORM ON MARRIAGE

	Age first marriage (1)	Number of first marriages per woman in each cohort (2)	Number marriages per woman in each cohort (3)
Treated	-0.056***	0.444	0.384
	(0.015)	(0.232)	(0.225)
	[0.005]	[0.164]	[0.224]
Treated*Post Reform	0.047*	-1.293**	-1.264**
	(0.023)	(0.217)	(0.211)
	[0.072]	[0.011]	[0.012]
Observations	2,322,360	9,106	9,118
R^2	0.051	0.372	0.387
BirthYear FE	YES	YES	YES
Region FE	YES	YES	YES
CalendarYear FE	NO	YES	YES
Mean pre-reform	24.81	29.58	30.33
Std. dev. pre-reform	4.807	26.31	26.01

Notes: The dependent variables are (1) the age of the women they married for the first time, (2) the percentage of (treated and control) women that married at least one time (multiplied by 1,000) and, (3) the total number of marriages divided by the total number of women born in each cohort (multiplied by 1,000). Regressions include cohort time and region fixed effects and (2-3) calendar year dummies. Note that we cannot include calendar year dummies when the dependant variable is age as it takes out all the variation. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Marriage registries (1976-2012), all women from cohorts 1961-1971.

Summing up, we conclude that the reform postponed first marriages and, consequently, postponed the age at which women had their children. Moreover, we find that the postponement in fertility is not away from the teenage years (before age 18), as the majority of previous literature has found; instead, our results show that the reform decreased the probability of women having the first child between the ages of 29 and 33. The reform increased the incidence of first births after the age of 34, which is an age at which women's fertility begins to drop, resulting in a reduction in completed fertility rates.[21]

3.3.2 Effect of the Reform on Infant Health at Delivery

We next focus on the potential long-term impacts of the reform. More precisely, we study whether the health of children born from women affected by the reform changed after the new policy was implemented. We measure children's health at the moment of delivery. If we find evidence that the reform has an impact on infant health, we can argue that child labor regulations can have intergenerational externalities that should be taken into account when thinking about the design of these policies.

In this analysis, we again use birth register data. We use four measures of newborn health: birth weight (in grams), the fraction of babies born weighing under 2,500 grams,[22] the fraction that are born after more than 37 gestational weeks,[23] and the fraction that die within the first 24 hours of life. Birth weight and survival of the first 24 hours data are only available from 1980 to 2014. Thus, when analyzing these outcomes, we drop the 1961 cohort from the pre-reform group and restrict the sample to all births that took place when the mother was between the ages of 18 and 43.[24]. It should also be noted that the birth weight is missing from 11 percent of all registered first births. However, as it can be observed from Table 3.A1 that the probability of not having registered birth weight is not affected by the reform.

We also investigate whether there is selection in the children that are actually born. It could be the case that, before the reform, those women born at the

[21] The fact that women's fertility rates decrease with age is well established, particularly for women over the age of 35. For instance, Leridon (2004) shows that the probability of conceiving after one year of trying decreases from 75 percent at age 30 to 66 percent at age 35.

[22] Babies born with less than 2,500 grams are considered to be of low birth weight by medical standards.

[23] We select 37 gestational weeks as a threshold because babies born earlier than that are medically considered premature.

[24] Note that we already showed in Table 3.4 that the reform did not have an effect on the probability of women having the first child before the age of 18 (from 14 and 18), so we are confident that we don't have a selected sample.

beginning of the year engaged in more unhealthy behaviors during pregnancy, which could lead to more fetal deaths. Then, the children that we observe from the women that were born at the beginning of the year would be those that come from the "better" mothers. To check this alternative channel, we use register data on late fetal deaths, which reports all natural abortions that took place when the fetus has at least six months of gestation. We do not find any significant differences between treatment and control women before and after the reform on the probability of suffering a premature fetus death of more than six months of gestation. However, medical research indicates that the greatest risk of suffering a natural abortion is during the first three months of gestation. Therefore, we cannot completely rule out the selection hypothesis with these results. Thus, we will also analyze the effect of the reform over the sex ratio (the probability of having a male first birth). This outcome can be considered a proxy for miscarriage as male births are known to miscarry more often. The medical literature argues that hormones induced by stress increase the probability of spontaneous abortions at an early stage of pregnancy, and these hormones have a larger effect on male than on female fetuses (Hobel et al., 1999; Byrne et al., 1987).

Note that this analysis only examines health at birth of the woman's first child. We include this restriction because a poor health outcome from the first birth can influence the decision to have a second child, as pointed out by Wolpin (1993).

Table 3.5 reports the effects of the reform on the sex ratio and the four infant health outcomes using the same econometric model as before. First of all, we find that the reform did not have any effect on the sex-ratio. After the reform, treated women did not have more first-born children of a certain gender, which provides further evidence that differential miscarriage is not a possible problem in our setting. Furthermore, we find that the reform has a negative impact on the health of children born to treated women (women born at the beginning of the year).[25] After the reform, the first child of a woman born at the beginning of the

[25] Results are robust in sign and significance to the substitution of cohort time dummies by linear, quadratic and quartic pre- and post-reform trends, the exclusion of region fixed effects, and the incorporation of the interactions of cohort and region fixed effects.

year has a 0.223 percentage-point (0.24%) higher probability of being premature (born with less than 37 gestational weeks). This translates into 290 more children of women of the first post-reform cohort (born in 1967) that are premature due to the reform.[26]

Table 3.5: EFFECT OF THE REFORM ON INFANT HEALTH OUTCOMES

	Infant health					
	Prob. male (1)	Maturity (2)	Multiple birth (3)	Survival 24h (4)	Weight (5)	Weight less 2,500 (6)
Treated	0.214	0.062	0.017	0.012	2.016	-0.038
	(0.284)	(0.037)	(0.043)	(0.015)	(1.813)	(0.052)
	[0.563]	[0.157]	[0.827]	[0.502]	[0.252]	[0.427]
Treated* Post Reform	-0.484	-0.223***	0.239**	-0.019	-4.694*	0.186**
	(0.613)	(0.049)	(0.079)	(0.015)	(2.035)	(0.060)
	[0.469]	[0.001]	[0.023]	[0.294]	[0.099]	[0.025]
Observations	2,493,107	2,493,107	2,493,107	2,173,324	1,938,272	1,938,272
R^2	0.000	0.013	0.021	0.000	0.011	0.009
BirthYear FE	YES	YES	YES	YES	YES	YES
CalendarYear FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Mean pre-reform	0.904	90.43	2.518	99.77	3218	6.466
Std. dev. pre-reform	0.294	29.42	15.67	4.794	506.1	24.59

Notes: The dependent variables are (1) the probability that the first birth is a boy, (2) the probability of having a first child 37 weeks of gestation, (3) the probability of having multiple births, (4) the probability of having a first child that survives the first 24 hours after delivery, (5) the weight at birth of the woman's first child and, (6) the probability that the first child is born with less than 2,500 grams. Regressions include cohort, calendar time and region dummies. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1971. For birth-weight, only consider the birth registries from 1980-2014 and cohorts of women 1962-1971.

Apart from the infant health outcomes, we also find that the reform increased the probability of having a multiple birth in 0.239 percentage points. This might be a consequence of the postponement of the entrance into motherhood. As shown before, we find that the reform increased the incidence of first births after the age

[26] 2,789 if we take into account the 10 consequent cohorts.

of 35, the age when women's fertility begins to drop. Many of these women might start receiving infertility treatments, which have a higher probability of multiple pregnancy. Also, at ages 35 or more, the probability of having multiple births increases, even without fertility treatments.[27] On the other hand, we do not find any effect of the reform on the probability of the first child surviving the first 24 hours.

The reform also caused women born at the beginning of the year to have children that weighted 4.69 grams less, on average, compared to children of women born at the end of the year. While 4.69 grams may not seem like a lot, it has to be taken into account that this is the estimated average impact of the reform. In fact, this result is of similar magnitude as the change in birth weight brought on by several US federal nutrition programs. For instance, Hoynes et al. (2011) determine that the Supplemental Program for Women, Infants and Children in the United States led to an increase in average birth weight of around 2 grams. Similarly, Almond et al. (2011) estimate that the US Food Stamp program increased the average birth weight between 2 and 5 grams. Moreover, we estimate that after the reform, women born at the beginning of the year have a 0.186 percentage points higher probability of having a first child with a low birth weight (less 2,500 grams). In absolute numbers, this implies that 453 more children are born with low weight from the cohort of women born in 1967.[28] As the percentage of children born weighing less than 2,500 grams is not very large,[29] this effect implies an increase of 2.7 percent in the number of low birth weight children due to the reform. These numbers constitute an important impact of the reform, as the long-run negative outcomes associated with low birth weight, such as labor market earnings and education, have been widely established in the literature (see Black et al. (2005), Figlio et al. (2014), Cook and Fletcher (2015) or Behrman and Rosenzweig (2004) , for instance).

[27] Given that the reform affects the probability of having multiple births, we cannot examine the effects of the reform on infant health outcomes excluding multiple births, as the resulting first births will constitute a selected sample. Though, if we restrict the sample to single first births, the effects of the reform over infant health go in the same direction.

[28] 4,352 in the subsequent 10 generations of women.

[29] Only 6.4% of children from the cohort of women born before 1966 were born weighing less than 2,500 grams.

Our results conflict with the scarce evidence presented in the extant literature, which finds either a positive impact of maternal education on child health (Currie and Moretti, 2003), or no causal effect (McCrary and Royer, 2011). Thus, in the next subsection, we propose three possible channels through which the child labor reform could have a negative impact on infant health.[30]

Explanatory Mechanisms

The postponement of first births

A first channel through which the reform operates is the postponement of the entrance into motherhood. We have shown in Table 3.3 that the reform increases the probability that women have their first child after the age of 34. Previous medical literature has indicated that having a first birth after the age of 35 could have negative effects on infant health as risk during pregnancy increases after that age. For instance, Jolly et al. (2000) find that advanced maternal age is correlated with an increased likelihood of delivering a small for gestational age baby, which may be related to poorer placental perfusion or transplacental flux of nutrients. Likewise, older women are more likely to deliver preterm.

Changes in the maternal marital status

We argued before that the reduction and postponement of fertility may be the result of a similar postponement and reduction of marriage. In this section, we focus on the change in the marital status of the subgroup of women that decide to become mothers. Previous literature (Gaudino et al., 1999; Bennett, 1992; Balayla et al., 2011) has established that children whose mothers are not married or have no father in the birth certificate data have worse health outcomes at the time of delivery. Table 3.6 shows that the reform significantly increased the probability that first children did not have a registered father by 0.219 percentage points (0.2%) and the probability that the mother is not married by 0.289 percentage points (0.33%). Therefore, a second possible mechanism through which the reform could be detrimental for infant health is the increase in the number of unmarried mothers and children without father.

[30] We are aware that the three channels we report in this paper might not be the only possible channels for the effect of the reform on infant health.

Table 3.6: EFFECT OF THE REFORM ON MARITAL STATUS OF MOTHERS

	Has father (1)	Mother married (2)
Treated	0.271*** (0.043) [0.002]	0.315** (0.121) [0.040]
Treated* Post Reform	-0.219*** (0.040) [0.001]	-0.289** (0.111) [0.041]
Observations	2,493,107	2,493,107
R^2	0.029	0.044
BirthYear FE	YES	YES
CalendarYear FE	YES	YES
Region FE	YES	YES
Mean pre-reform	96.93	88.86
Std. dev. pre-reform	17.25	31.46

Notes: The dependent variables are (1) the probability that the child has a father, and (2) the probability that the mother is married. Regressions include cohort, calendar year, and region dummies. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth registries (1980-2014), all women from cohorts 1961-1971.

Changes in labor market behavior and health habits

A third channel through which the reform could be affecting infant health is through changes in labor market behavior. It seems plausible that, if the child labor reform proofs to have increased the educational attainment of treated women, it could also have affected their probability of working or the type of job that they have. Thus, the level of stress of treated women or even their health behaviors could be changed due to the changes in labor market outcomes (through, for example, an income effect). This, in turn, could affect the health of their babies at birth. In fact, previous literature has demonstrated the association between increased education and the prevalence of unhealthy behaviors (especially smoking) among Spanish women, converging toward men's behaviors (see

Pampel (2003), and Schiaffino et al. (2003), for the Spanish case).

To examine the potential long-term labor market effects of the reform on affected women at ages 34-56 we use the Labor Force Survey (LFS), which provides labor market information from 2000 to 2013. Table 3.7 shows that the reform has a positive but not significant impact on the probability of working. We also observe that, after the reform, treated women have a lower probability of being in a low skill job or having a part-time job as compared to women born at the end of the year. Both results suggest potentially more demanding jobs which could be linked to potentially higher stress as well as income levels.

Table 3.7: Effect of the Reform on Labor Outcomes of Women

	Work	Low skill job	Part-time job
	(1)	(2)	(3)
Treated	-0.580	1.053**	0.727**
	(0.324)	(0.378)	(0.319)
	[0.112]	[0.020]	[0.046]
Treated*Post Reform	0.508	-1.010*	-1.260***
	(0.402)	(0.473)	(0.368)
	[0.244]	[0.056]	[0.005]
Observations	151,602	92,352	58,349
R^2	0.035	0.015	0.008
BirthYear FE	YES	YES	YES
Region FE	YES	YES	YES
CalendarYear FE	YES	YES	YES
Mean pre-reform	59.02	25.03	20.28
Std. dev. pre-reform	49.18	43.32	40.21

Notes: The dependent variables are (1) the probability of working at the time of the survey, (2) the probability of having a low skill job and, (3) the probability of having a part-time job. Regressions include cohort, calendar year, and region dummies. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Spanish Labor Force Survey (2000- 2013), all women from cohorts 1958-1974.

Finally, we use data from the Spanish National Health Survey (see the data appendix for more information on this database) to determine whether the labor market impacts of the child labor reform are also translated into differences in health behaviors of treated women after the reform.[31] Table 3.8 shows that, after the reform, women born at the beginning of the year have a higher probability of smoking regularly and they also smoke more cigarettes per day.[32] Although the reform did not impact the probability that these women are ex-smokers, we do find that after the reform women born at the beginning of the year have a lower probability of quitting smoking during pregnancy (conditional on being smokers and having kids). These outcomes could directly affect the health of their offspring.

This unexpected result is likely due to these cohorts growing up during the early post-Franco era. Women in these cohorts experienced the process of gender equalization, continuously getting more education and increasing their participation in the labor market. Consequently, access to and social acceptance of smoking were much higher than for previous (pre-reform) cohorts. For instance, a recent paper by Bilal et al. (2015) shows a high negative correlation between gender inequality and the female-to-male smoking ratio in Spain from the 1960s to the 2010s.

More importantly, this positive association between education and prevalence of smoking for women cannot be considered an isolated and particular case of Spain at that time. In many countries in the world, the number of smoking women is increasing, even though smoking prevalence among women is still lower than among men. This phenomenon can be attributed to the weakening of the social and cultural constraints that prevented many women from smoking in the past (Mackay and Amos, 2003). In some Eastern European countries and Eastern

[31] Although this survey is available for several years, only the 2006 wave reports the individual's month of birth, which is a crucial variable for our identification strategy. Therefore, the results that we report are for the 2006 wave and include cohorts from 1961 to 1971.

[32] Results are robust in sign and significance to the substitution of cohort time dummies by linear, quadratic and quartic pre- and post-reform trends, the exclusion of region fixed effects, and the incorporation of the interactions of cohort and region fixed effects.

Table 3.8: EFFECT OF THE REFORM ON BEHAVIORAL AND HEALTH OUTCOMES OF WOMEN

	Smoke/day (1)	Smoke regular (2)	Ex-smoker (3)	Pregnancy as motive for being ex-smoker (4)
Treated	-0.053*	-0.063**	0.007	0.058*
	(0.017)	(0.011)	(0.011)	(0.014)
	[0.068]	[0.013]	[0.570]	[0.086]
Treated*Post Reform	0.115***	0.111***	-0.034	-0.079**
	(0.024)	(0.031)	(0.020)	(0.029)
	[0.002]	[0.003]	[0.158]	[0.023]
Observations	3,151	3,151	3,151	2,269
R^2	0.018	0.019	0.014	0.024
BirthYear FE	YES	YES	YES	YES
Region FE	YES	YES	YES	YES
Mean pre-reform	0.330	0.362	0.235	0.0481
Std. dev. pre-reform	0.471	0.481	0.424	0.214

Notes: The dependent variables are (1) the probability of smoking at least one cigarette a day, (2) probability of smoking regularly, (3) the probability of having quitted smoking and (4) the probability of having quitted smoking during pregnancy, conditional on being an ex-smoker. The regression include cohort and region dummies. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Spanish National Health Survey (2006), all women from cohorts 1961-1971.

Mediterranean countries a high smoking prevalence among high educated women compared to low educated women has been established by previous literature (Bosdriesz et al., 2014). This same pattern has been found to hold (Pampel, 2003) in other high-income countries at early stages of the smoking epidemic. Then, the process of gender equalization and the initial adoption of tobacco consumption that was taking place in the early post-Franco era in Spain could explain the positive correlation that we find between smoking prevalence and education among women. Those women affected by the child labor reform had higher education and financial independence that improved their social status and hence an autonomy to emulate their male counterparts' life style.

Thus, we find evidence that the child labor reform has positive impacts on affected women, as they have a lower probability of being in low skill jobs, a lower probability of having a part-time job and they also increase their educational attainment. However, the impact of the reform on their children is negative. The better employment perspectives of women also increases the women's probability of engaging in unhealthy behaviors that result in poorer health outcomes for their first child at the moment of delivery. We have also show that part of these negative health outcomes on children can be attributed to the postponement of fertility after the age of 35 and the increase in unmarried mothers and children without fathers.

3.4 Robustness Checks

In this section, we perform several robustness checks of our key results. More specifically, we explore the sensitivity of our results to the exclusion of some pre-reform cohorts that could be considered non-compliers and perform some placebo tests in which we change the definition of the timing of the reform (fake reforms). We also examine the influence of father's education on fertility and infant health outcomes by including additional variables that control for the characteristics of the father. Finally, we analyze the robustness of our results when we reduced the sample to include only individuals born during the middle months of the year.

3.4.1 Exclusion of Possible Non-Compliers

The child labor reform took place in 1980, affecting cohorts of individuals under the age of 14 that year. This means that individuals born from 1966 onwards are affected by the reform. In our main analysis we drop the 1966 cohort because it included individuals that turned 14 the year of the reform and we found it difficult to predict the effect of the reform on this cohort. However, the cohorts of 1965 and 1964 were 15 and 16 years old the year the reform was implemented and were, therefore, likely to be partially affected by it. The 1980 law was unclear about the consequences for individuals who were already working at the age of 14 and 15 when the reform was introduced. Thus, we perform the same analysis but additionally dropping the cohorts of 1964 and 1965, as they could potentially represent non-compliers.

The results in Table 3.9, 3.10, 3.11 and indicate that the effects of the reform on fertility and infant health outcomes are unchanged when we exclude these two additional cohorts. Thus, we conclude that our results are robust to the exclusion of possible non-compliers.

3.4.2 Placebos

We also perform several placebo tests in which we use "fake" reform years. In these tests we only include those cohorts of women not affected by the "real" reform (the reform in 1980). We examine the effect of three "fake" reforms affecting the cohorts of 1961, 1962, and 1963.[33] We use the same econometric specification and treatment status definition as before. We expect a nonsignificant effect of the interaction term between the post-reform dummy and the treatment dummy.

In Figure 3.3 and 3.4 , we plot the estimates of the interaction term and the 95 percent confidence interval for the different fertility and infant health outcomes. Graph a) of Figure 3.3 shows that none of the "fake" reforms considered has a significant effect on the age at which women have their first child. Moreover, graphs d) and e) of Figure 3.3 again indicate no effect of any of the "fake" reforms on the probability of having a child or the total number of children that each woman has.

We perform the same analysis for all four infant health outcomes. We see from graphs b) and e) in Figure 3.4 that the "fake" reforms for the 1962, 1963, and 1964 cohorts do not affect the probability of having a first child with more than 37 weeks of gestation or the child's birth weight. The results on multiple

[33] We cannot replicate the placebo tests for the cohorts of 1964 and 1965 because, as explained above, they are potentially partially influenced by the reform. These two cohorts were 15 and 16 years old when the reform was introduced. Thus, if they were not working at that moment, the reform would have prevented them from start working. Moreover, these cohorts could also still be in the last year of primary schooling if they had to retake a year at school.

births and survival during the first 24 hours are less clear, as the trend difference between the treatment and control groups seems to change for some cohort.

In sum, we believe that the placebo tests provide us with reasonable evidence to argue that there are no significant trend changes among the treatment and control groups for the cohort of women not affected by the reform for the majority of the fertility and infant health outcomes considered.

3.4.3 Influence of Father's Education on Fertility and Infant Health Outcomes

It is reasonable to think that couples make fertility decisions jointly. If this is the case, many of our results related to fertility and infant health outcomes may not only be driven by the effect of the child labor reform on the mother but also by its effects on the child's father. In this section, we examine whether the effect of maternal education on fertility and infant health outcomes hold up when controlling for paternal education. We proxy education of the father by the average age at which men have their first child, calculated by cohort, region, and treatment status. We also control for the probability that children have a registered father, similarly calculated by cohort, region, and treatment status.

We can indirectly check the relevance of this instrumental variable by analyzing the effect of the reform on the age when fathers had their first child. The first regression in Table 3.12 shows that the reform increased the age at which fathers had their first child by almost a month, indicating that our instrument is relevant.

Table 3.12 also shows that when controlling for the fathers' characteristics, the effect of the reform on the fertility outcomes is considerably reduced. This confirms our hypothesis that not only the effect of the reform over treated women but also over treated men is affecting the age at which women have their first child. On the contrary,, the effect of the reform on infant health outcomes is quite robust

Table 3.9: Robustness Check: Effect of the Reform on Fertility Outcomes Excluding Possible Noncompliers

	Drop cohort 1966 (1)	Drop cohorts 1966-65 (2)	Drop cohorts 1966-65-64 (3)
	Age when first child		
Treated	-0.080**	-0.089***	-0.094***
	(0.009)	(0.005)	(0.002)
	[0.023]	[0.002]	[0.003]
Treated* Post Reform	0.058***	0.067***	0.071***
	(0.011)	(0.008)	(0.006)
	[0.002]	[0.002]	[0.003]
Observations	2,493,107	2,238,017	1,974,964
BirthYear FE	YES	YES	YES
Region FE	YES	YES	YES
CalendarYear FE	NO	NO	NO
	Perc. women in each cohort become a mother		
Treated	1.319**	1.358*	1.382**
	(0.203)	(0.228)	(0.258)
	[0.037]	[0.052]	[0.035]
Treated* Post Reform	-1.627***	-1.599***	-1.532***
	(0.123)	(0.138)	(0.155)
	[0.003]	[0.005]	[0.003]
Observations	9,982	8,983	7,983
BirthYear FE	YES	YES	YES
Region FE	YES	YES	YES
CalendarYear FE	YES	YES	YES
	Number of children per women in each cohort		
Treated	3.034***	3.182***	3.225***
	(0.390)	(0.445)	(0.497)
	[0.004]	[0.002]	[0.003]
Treated* Post Reform	-3.079***	-3.118**	-3.009***
	(0.187)	(0.234)	(0.281)
	[0.003]	[0.017]	[0.003]
Observations	10,026	9,023	8,018
BirthYear FE	YES	YES	YES
Region FE	YES	YES	YES
CalendarYear FE	YES	YES	YES

Notes: The dependent variables are (Panel 1) the age at which women had their first child, (Panel 2) the percentage of women in each cohort that had at least one children, and (Panel 3) the total number of children per each cohort. Regressions include cohort, region dummies and (Panel 2 and 3) calendar year. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets.* significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1971.

Table 3.10: Robustness Check: Effect of the Reform on Infant Health Outcomes Excluding Possible Noncompliers

	Drop cohort 1966	Drop cohorts 1966-65	Drop cohorts 1966-65-64
	(1)	(2)	(3)
		Infant health: Prob. male	
Treated	0.214	-0.039	-0.240
	(0.284)	(0.201)	(0.117)
	[0.563]	[0.880]	[0.332]
Treated*Post Reform	-0.484	-0.230	-0.027
	(0.613)	(0.580)	(0.557)
	[0.469]	[0.712]	[0.949]
Observations	2,493,107	2,238,017	1,974,964
		Infant health: Maturity	
Treated	0.062	0.083	0.096
	(0.037)	(0.041)	(0.054)
	[0.157]	[0.130]	[0.253]
Treated*Post Reform	-0.223***	-0.239***	-0.246**
	(0.049)	(0.052)	(0.062)
	[0.001]	[0.002]	[0.011]
Observations	2,493,107	2,238,017	1,974,964
		Infant health: Multiple births	
Treated	0.017	0.020	0.011
	(0.043)	(0.054)	(0.072)
	[0.827]	[0.849]	[0.996]
Treated*Post Reform	0.239**	0.234*	0.238
	(0.079)	(0.085)	(0.097)
	[0.023]	[0.060]	[0.121]
Observations	2,493,107	2,238,017	1,974,964
BirthYear FE	YES	YES	YES
Region FE	YES	YES	YES
CalendarYear FE	YES	YES	YES

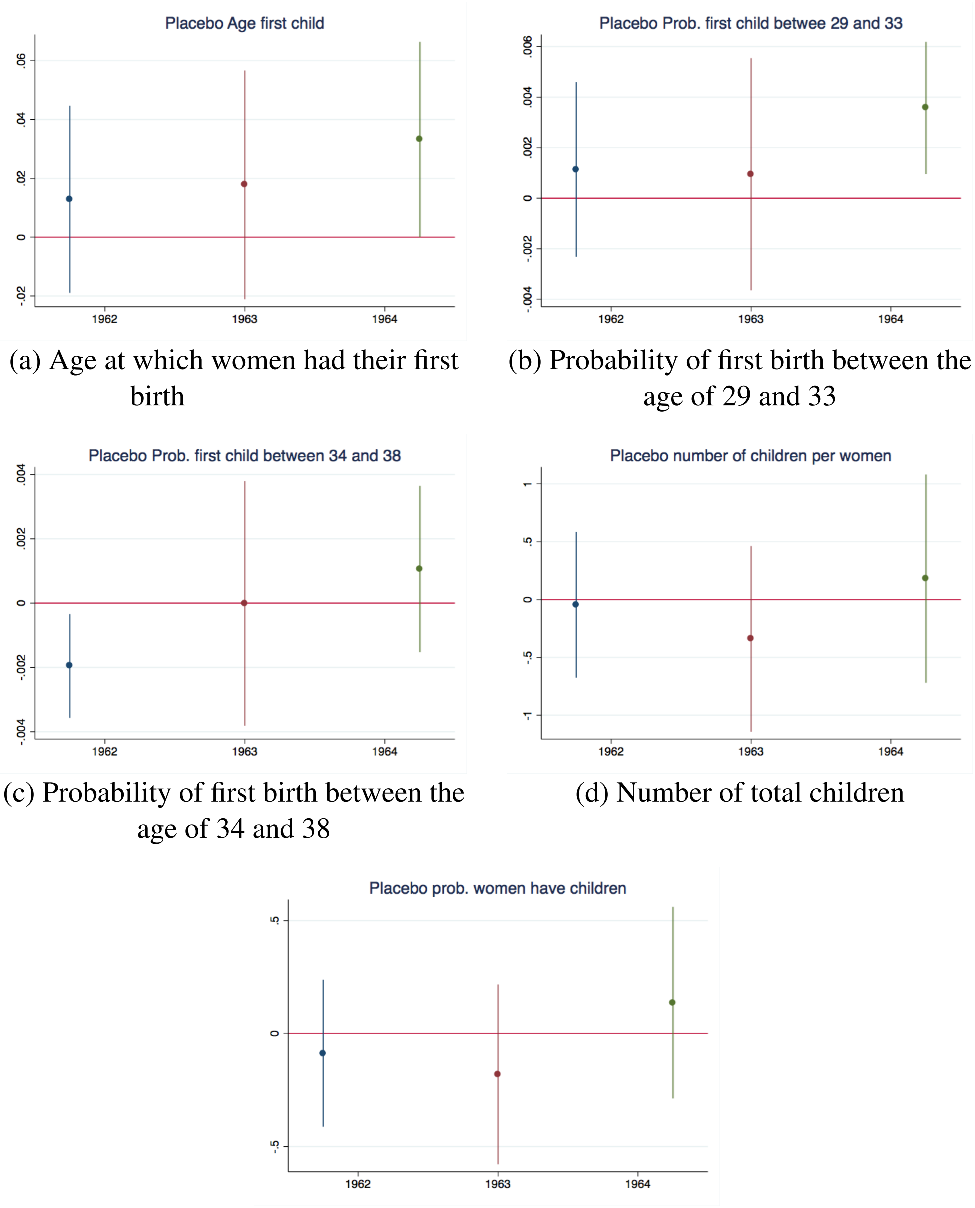
Notes: The dependent variables are (Panel 1) the probability of having a first birth boy, (Panel 2) the probability of having a first child with equal or more than 37 weeks of gestation, and (Panel 3) the probability of having a multiple first births. Regressions include cohort, calendar time, and region dummies. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets.* significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1971. For birth-weight, only consider the birth registries from 1980-2011 and cohorts of women 1962-1971.

Table 3.11: ROBUSTNESS CHECK: EFFECT OF THE REFORM ON INFANT HEALTH OUTCOMES EXCLUDING POSSIBLE NONCOMPLIERS (CONT.)

	Drop cohort 1966	Drop cohorts 1966-65	Drop cohorts 1966-65-64
	(1)	(2)	(3)
		Infant health: survival 24 h	
Treated	0.012	0.020	0.039
	(0.015)	(0.018)	(0.008)
	[0.502]	[0.253]	[0.117]
Treated*Post Reform	-0.019	-0.027	-0.047***
	(0.015)	(0.018)	(0.009)
	[0.294]	[0.261]	[0.007]
Observations	2,173,324	1,927,458	1,674,312
		Infant health: birth weight	
Treated	2.016	4.020**	5.129
	(1.813)	(0.866)	(0.434)
	[0.252]	[0.011]	[0.382]
Treated*Post Reform	-4.694*	-6.585***	-7.534**
	(2.035)	(1.206)	(0.884)
	[0.099]	[0.003]	[0.023]
Observations	1,938,272	1,721,889	1,502,078
		Infant health: less 2500 grams at birth	
Treated	-0.038	-0.092***	-0.136
	(0.052)	(0.035)	(0.018)
	[0.427]	[0.003]	[0.320]
Treated*Post Reform	0.186**	0.235**	0.271*
	(0.060)	(0.044)	(0.030)
	[0.025]	[0.011]	[0.054]
Observations	1,938,272	1,721,889	1,502,078
BirthYear FE	YES	YES	YES
Region FE	YES	YES	YES
CalendarYear FE	YES	YES	YES

Notes: The dependent variables are (Panel 1) the probability of having a first child that survives the first 24 hours after delivery, (Panel 2) birth weight, and (Panel 3) the probability that the first child is born with less than 2,500 grams. Regressions include cohort, calendar time, and region dummies. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets.* significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1971. For birth-weight, only consider the birth registries from 1980-2011 and cohorts of women 1962-1971.

Figure 3.3: PLACEBOS ON FERTILITY



(a) Age at which women had their first birth

(b) Probability of first birth between the age of 29 and 33

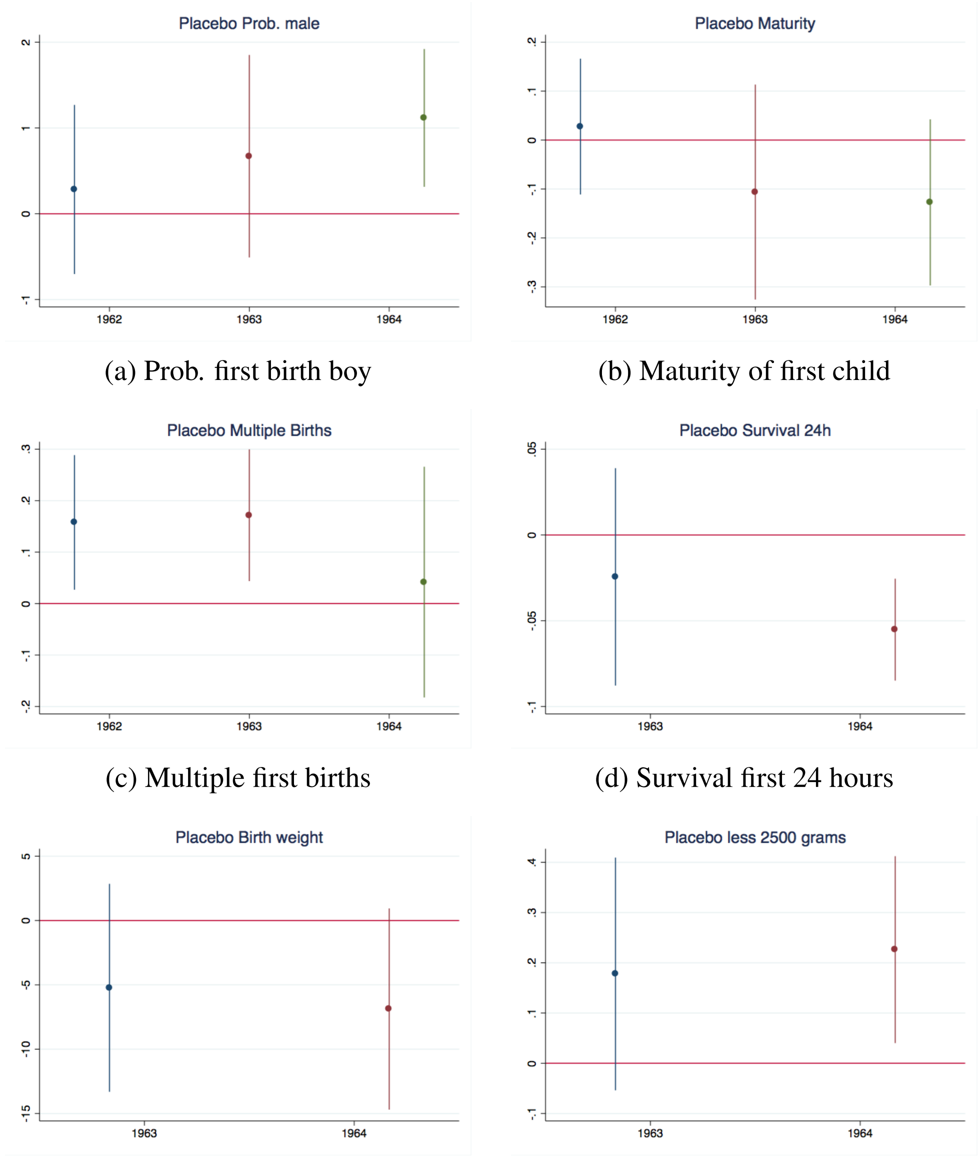
(c) Probability of first birth between the age of 34 and 38

(d) Number of total children

(e) Probability of having children

Notes: We report the point estimates and the 95% confidence interval of the interaction term of the treatment and the ”fake” reform taking place for the cohorts of 1962, 1963, and 1964. We only consider cohorts not affected by the real reform: 1961-1965. The *treatment* is defined as those women born from January to June and *control* those women born from July to December. Robust standard errors clustered at cohort level. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1965 that had a child for the first time.

Figure 3.4: Placebos on Infant Health



(a) Prob. first birth boy

(b) Maturity of first child

(c) Multiple first births

(d) Survival first 24 hours

(e) Birth-weight of first child

(f) Probability of weighting less 2,500 grams

Notes: We report the point estimates and the 95% confidence interval of the interaction term of the treatment and the ”fake” reform taking place for the cohorts of 1962, 1963, and 1964. We only consider cohorts not affected by the real reform: 1961-1965. The *treatment* is defined as those women born from January to June and *control* those women born from July to December. Robust standard errors clustered at cohort level. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1965 that had a child for the first time.

to the inclusion of controls for paternal characteristics. In addition, the instruments for fathers' characteristics are not significantly associated with any of the infant health outcomes (except for maturity and multiple births). This last result implies that the mother's characteristics could be a more important determinant of infant health at the time of delivery, reinforcing our finding that part of the negative infant health outcomes are attributable to the effect of the reform on women's health behaviors, such as smoking.

3.4.4 Reduced Sample

Previous literature has pointed out that individuals born at the beginning of the year are typically quite different in several dimensions from individuals born at the end of the year (Bound and Jaeger, 2000; Buckles and Hungerman, 2013). Note, however, that this is not a necessary assumption in our identification strategy to correctly estimate the causal effect of the reform. The identifying assumption needed in our approach is that any difference that we observe for women born at the begin and end of the year after the reform (with respect to the differences observed before the reform) is due to the passing of the child labor reform.

Thus, even though the existing differences among women in the pre-reforms cohorts should not affect our results, we try to address the remaining possible doubts by omitting from our sample those individuals born in the first and last two months of each year.[34] Therefore, we analyze the robustness of our main results on fertility and infant health outcomes comparing only women born in months 3, 4 and 5 with women born in months 8, 9 and 10 before and after the reform. Table 3.13 shows that even with this reduced sample approach our main findings of the effect of the reform on the fertility and infant health outcomes are unchanged.

[34] Another robustness check (available upon request) shows that the impact of the reform is the same for individuals born during January, February, March, April or May. This results confirms that the effect of the reform comes from completing primary school and not from adding months of schooling.

Table 3.12: Effect of the Reform on Fertility and Infant Health Outcomes Controlling for the Father

			Infant Health					
	Age men	Age women			Multiple			
	first birhts	first births	Prob. male	Maturity	births	Survival 24h	Weight	Weight less 2,500
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated	-0.075*	0.067***	0.064	-0.171	-0.042	0.013	2.874	-0.051
	(0.015)	(0.011)	(0.353)	(0.060)	(0.049)	(0.017)	(2.302)	(0.077)
	[0.059]	[0.003]	[0.890]	[0.104]	[0.428]	[0.502]	[0.345]	[0.576]
Treated* Post Reform	0.084**	0.023*	-0.459	-0.170**	0.259**	-0.019	-5.079*	0.161*
	(0.026)	(0.012)	(0.623)	(0.049)	(0.079)	(0.015)	(2.079)	(0.077)
	[0.011]	[0.091]	[0.495]	[0.014]	[0.017]	[0.314]	[0.091]	[0.087]
Aver. age men		0.897***	-1.115	-1.471***	-0.262**	0.012	2.011	-0.177
		(0.022)	(1.112)	(0.342)	(0.115)	(0.032)	(7.252)	(0.277)
Perc. fathers		-15.423***	49.180	33.124*	-10.116	-1.330	392.286	-4.417
		(1.353)	(57.352)	(16.894)	(6.082)	(1.665)	(289.494)	(11.358)
Observations	2,250,019	2,493,107	2,493,107	2,493,107	2,493,107	2,173,324	1,938,272	1,938,272
R^2	0.057	0.068	0.000	0.013	0.021	0.000	0.011	0.007
BirthYear FE	YES	YES	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES	YES	YES
CalendarYear FE	NO	NO	YES	YES	YES	YES	YES	YES
Mean pre-reform	29.70	26.83	289.4	90.43	2.518	99.77	3218	7.425
Std. dev. pre-reform	5.359	5.507	299.8	29.42	15.67	4.794	506.1	26.22

Notes: The dependent variables are (1)the age of the men when they had their first child, (2) the age of the women when they had their first child, (3) the probability of having a first birth male, (4) the probability of having a first child with 37 weeks or more of gestation, (5) the probability of having multiple first births, (6) the probability of having a first child that survives the first 24 hours after delivery, (7) birth-weight at the time of delivery, and (8) the probability that the first child is born with less than 2,500 grams. Regressions (1-8) include cohort, (3-8) calendar time and (1-8) region dummies. We also control for the average age of the father of the first birth and for the probability that the first child has a father. These two variables are calculated by cohort-region and treatment. *Treated* are individuals born from January to May and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth registries (1975-2014), all individuals from cohorts 1961-1971. For birthweight and survival 24h, we only consider the birth registries from 1980-2011 and cohorts of women 1962-1971.

Table 3.13: Effect of the Reform on Fertility and Infant Health Outcomes with a Reduced Sample

		Fertility	
	Age first birth (1)	Perc. women in each cohort become a mother (2)	Number of children per women in each cohort (3)
Treated	-0.086*** (0.015) [0.001]	0.838*** (0.266) [0.001]	2.342*** (0.525) [0.001]
Treated* Post Reform	0.068** (0.021) [0.007]	-1.253*** (0.252) [0.001]	-2.706*** (0.477) [0.001]
Observations	1,383,799	9,883	9,939
BirthYear FE	YES	YES	YES
Region FE	YES	YES	YES
CalendarYear FE	NO	YES	YES
		Infant health	
	Prob. male (4)	Maturity (5)	Multiple birth (6)
Treated	0.260 (0.451) [0.529]	0.047 (0.038) [0.283]	0.031 (0.069) [0.725]
Treated* Post Reform	-0.836 (0.747) [0.311]	-0.199*** (0.050) [0.005]	0.197* (0.102) [0.087]
Observations	1,383,799	1,383,799	1,383,799
	Survival 24h (7)	Weight (8)	Weight less 2500 (9)
Treated	0.018 (0.025) [0.502]	1.571 (1.437) [0.388]	-0.033 (0.082) [0.654]
Treated* Post Reform	-0.023 (0.027) [0.435]	-3.866** (1.524) [0.029]	0.176 (0.101) [0.162]
Observations	1,207,828	1,077,377	1,077,377
BirthYear FE	YES	YES	YES
Region FE	YES	YES	YES
CalendarYear FE	YES	YES	YES

Notes: The dependent variables are (1) the age of the women they had their first child, (2) the percentage of (treated and control) women that had at least one child (multiplied by 1,000), and (3) the total number of children divided by the total number of women born in each cohort (multiplied by 1,000), (4) the probability that the first birth is a boy (multiplied by 100), (5) the probability of having a first child 37 weeks of gestation (multiplied by 100), (6) the probability of having multiple births (multiplied by 100), (7) the probability of having a first child that survives the first 24 hours after delivery (multiplied by 100), (8) the weight at birth of the woman's first child, and (9) the probability that the first child is born with less than 2,500 grams (multiplied by 100). Regressions (1-9) include cohort, (2-9) calendar time and (1-9) region dummies. *Treated* are individuals born from January to May and *control* are those born from July to December. Robust standard errors clustered at cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth registries (1975-2014), all individuals from cohorts 1961-1971. For birthweight and survival 24h, we only consider the birth registries from 1980-2011 and cohorts of women 1962-1971.

3.5 Discussion

This study investigates the effect of a child labor regulation on fertility and infant health outcomes at the time of delivery. We exploit a reform implemented in Spain in 1980 that increased the minimum legal working age from 14 to 16 years old. Before the reform, students born at the beginning of the year had different incentives to finish primary education than those born at the end of the year. The introduction of the reform abolished these different incentives. Thus, we exploit the within-cohort variation following a difference-in-differences approach by comparing individuals born during the first or last six months of the year, before and after the reform.

Jiménez-Martín et al. (2015) showed that the reform was enforced and was effective. Those women and men born at the beginning of the year (that had lower educational attainment before the reform) had higher incentives to finish primary education and continue secondary and post-secondary education after the reform.

We find that, as a consequence, the reform prompted a postponement of first births by 21 days, on average. This number is very similar to the results in the majority of the previous literature studying a different type of reform that also increased educational attainment (compulsory schooling laws). However, our results show that this postponement is not followed by a catching-up effect, as the reform increased a woman's probability of ending her fertile lifecycle without any children and reduced her completed fertility. We find that after the reform 2,198 women born between 1967 to 1976 do not become mothers. In turn, this resulted in 4,160 fewer children born from the 1967–1976 cohorts of women.

We provide evidence that the lack of catching-up effect and the reduction in completed fertility operate through a postponement of first births until an age when the catching-up is more difficult. In fact, we show that the reform decreased the probability of pregnancy during the early thirties while increasing the probability of having late first births (after the age of 34).

The marriage market is another factor that contributes to the postponement of first births. We find that the reform increased the age at which women marry for the first time by almost half a month. This postponement of marriage also leads to a decrease in the likelihood of getting married and the total number of marriages per woman.

Finally, we focus on the effects of the reform on children's health at the moment of delivery. We find that, for mothers born at the beginning of the year, the reform increased the probability of having a first child at less than 37 gestational weeks by 0.223 percentage points. This result implies that women born between 1967 and 1976 had 2,789 more children born with less than 37 weeks of gestation. Moreover, these mothers also had a higher probability of having low birth weight babies after the reform.

We propose three different channels that could lead to this detrimental effect of the child labor law on children's health. The first is the effect of the increase of the age at which treated women get pregnant for the first time. When we control for the age at which women had their first child, the reform has a positive rather than a negative effect on infant health. This result is consistent with the positive impact of education on babies' health found in previous literature. Thus, we attribute our finding of negative impacts of the child labor law on infant's health to more educated mothers having their first child at an older age, making their pregnancies more risky and increasing the chances of poor infant health outcomes.

The second channel operates through changes in maternal marital status. We show that the reform increased the number of unmarried mothers as well as the number of children without fathers. Previous literature has proven that the lack of a father can be detrimental for the health of the baby at the moment of delivery.

The third channel that we propose is changes in labor market prospects and unhealthy habits of affected women. More precisely, we find that the reform decreased the probability of treated women having a low skill job or a part-time job. Simultaneously, their unhealthy habits increased as they increased their smoking prevalence. Thus, the fact that after the reform more educated women had better labor market outcomes has a negative impact on pregnancy through the increase in unhealthy behaviors. More precisely, we find that the probability of quitting smoking during pregnancy is reduced for women born at the beginning of the year after the reform.

Therefore, we conclude that even though the child labor reform had positive impacts on women by increasing their educational attainment and improving their labor market prospects, the reform had negative consequences for their children. This effect is driven by the increase in women's age at delivery, the increase of unmarried mothers and children without fathers, and by the increase in women's unhealthy habits. These results have to be considered together with the fact that the child labor reform that we are analyzing took place during the 1980s in Spain. At that time, Spain was still a developing country; a high percentage of its population had low levels of education and entered the labor market at an early age. Furthermore, as we show in this paper, the level of labor market integration and educational attainment among pre-reform women cohorts was very different from that of men. Thus, the results we find in this paper are more relevant, from a policy perspective, to developing countries whose educational system, child labor market participation rates, and social development are similar to the levels that Spain was experiencing around 1980.

3.6 Appendix

3.6.1 Pre-Post Analysis of the Reform

Our identification strategy relies on comparing individuals in the same cohort before and after the implementation of the policy. Because the reform took place during a time of social upheaval in Spain, we do not want to rely solely on before-and-after differences.[35] Thus, we employ a more conservative strategy, comparing women within the same cohorts. In this section, however, we provide some graphical evidence of the potential overall effect of the ET reform on some of the more important outcomes.

Figure 3.A2 shows that the reform decreased the probability of having the first child between the age of 24 and 28 by 2.6 percentage points and the probability of having the first child between the age of 29 and 33 by 0.9 percentage points.[36] However, the reform increased the probability of having the first child between the age of 34 and 38 and between the age of 39 and 43 by 1.3 and 0.1 percentage points respectively.

This postponement of motherhood is accompanied by a decrease in the number of women that become a mother by 0.16 for every 1,000 women and by a decrease of 0.28 children for every 1,000 women. The first two graphs of Figure A3 illustrate these effects. Finally, we also find that the reform has a negative impact on infant health. The probability of having a premature child increases by 0.36 percentage points, and the probability of having a multiple first births rises by 0.16 percentage points, as seen in the last graphs of Figure 3.A3.

[35] For instance, divorce was legalized in 1981 and abortion in 1985.

[36] In this estimation, we consider the cohorts of 1961–1965 to be the pre-reform cohorts, and those of 1966– 1971 to be the post-reform cohorts. We drop the cohort of 1966 in this analysis, as these women turned 14 the year the reform was introduced, 1980. The econometric model includes linear and quadratic trends, and clusters the standard errors at the cohort level.

3.6.2 Data Appendix

Throughout this paper we use different databases. In this section, we aim to describe these databases and explain the main variables used in our previous analysis.

Spanish Labor Force Survey

The Spanish Labor Force Survey is a continuous quarterly survey that contains information related to the labor market, active unemployment and inactivity of the population living in family dwelling in Spain. This database is available since 1964 however, in this paper, we use this database from 2000 to 2013 (for education outcomes) or from 2000 to 2007 (labor market outcomes), as the month of birth was not specified before. We will only consider in our sample women born between 1958 and 1974. We drop from our final sample all individuals not born in Spain and those individuals born in 1966 and therefore turned 14 the year the reform took place (1980). At the end we have information about 320,566 individuals from 2000 to 2013 and 180,573 from 2000 to 2007.

We use this data to assess the impact of the reform on education attainment (see Section 2) and labor market outcomes (see Section 3.2.1). For educational attainment, we use a variable that specifies the maximum level of education attained by all individuals with more than 16 year old at the moment of the interview, as well as, the self-reported age that they had when they acquired the maximum level of education. For the labor market outcomes we use a variable that ask for the employment situation the week before the interview of all individuals older than 16 at the moment of the interview, as well as, the type of occupation they have.

Therefore, the main variables used are the following and their descriptive statistics can be found in Table 3.A2 :

- **Early School Leaver**: A dummy that is equal to one if the individual is illiterate, have not completed the first eight years of education, or has been enrolled in labor market integration programs that do not require finishing the first eight year os education, or zero otherwise.

- **Drop with less or 16 year old**: A dummy that is equal to one if the individual has drop out of school before or at 16 years old. Note that education is only compulsory until 14 years old, or zero otherwise.

- **Work**: A dummy that is equal to one if the woman was working the week before to the interview, or zero otherwise

- **Inactivity**: A dummy that is equal to one if the woman was not participating in the labor market one week before the interview, or zero otherwise.

- **High skill job**: A dummy that is equal to one if the woman has a job that can be classified as business manager or administrator, civil servant, scientific and intellectual technician or professional, or as support technician or professional, and zero otherwise.

- **Low skill job**:A dummy that is equal to one if the woman has a job that can be classified as blue-collar in manufacturing factories, construction, minery, plant and machine operators and assemblers, or other unskilled jobs , and zero otherwise.

Birth Statistics

This database contains administrative data from birth certificates for the universe of children born in Spain between 1975 and 2014. The information is self-reported as it comes from the Statistical Birth Bulletins, that are filled out by the parents, relatives or persons so obligated by law to declared the childbirth.

The raw microdata contains 18,602,664 births. We, then, restrict the sample to births of Spanish women born between 1961 and 1971 that had 14 to 43 years old at the moment of delivery. We also drop births of women born in 1966 and who therefore turned 14 the year the reform took place (1980) and those of women born in July. Thus, finally we observe a total of 4,520,086 births or

2,493,107 first births in our sample.

We use this database to assess the impact of the reform on fertility (see Section 3.1) and infant health outcomes (See Section 3.2). Here we define the main fertility and infant health variables used throughout the paper whose descriptive statistics could be found in Table 3.A3:

- **Age women when first child**: Age of the women when they had their first child.
- **First birth between certain ages**: A dummy variable that is equal to one if the woman has her first child between that ages, and zero otherwise.
- **Prob. male**: A dummy variable that is equal to one if the first child of the woman is a male, and zero otherwise (multiplied by 100).
- **Maturity**: A dummy variable that is equal to one if the woman has her first child with 37 or more weeks of gestations, and zero otherwise (multiplied by 100).
- **Survive 24h**: A dummy variable that is equal to one if the woman has her first child that survives the first 24 hours after delivery, and zero otherwise (multiplied by 100).
- **Birth Weight**: Weight at birth of the woman's first birth (multiplied by 100).
- **Weight less 2,500**: A dummy variable that is equal to one if the woman's first child is born with less than 2,500 grams, and zero otherwise (multiplied by 100).
- **Number of first births by cohort and treatment**: Total number of first births for a cohort of women divided by the total number of women of that same cohort multiplied by 1000.
- **Number of total births by cohort and treatment**: Total number of births for a cohort of women divided by the total number of women of that cohort multiplied by 1000.

Population and Housing Census of 2011

This database surveys a representative sample of 5 percent of the population living in Spain in 2011 and collects information about some the persons, households, buildings and dwellings.

The raw microdata contains information about 4,107,465 families. We, then, restrict the sample to f Spanish women born between 1961 and 1971. We also drop women born in 1966 and who therefore turned 14 the year the reform took place (1980) and those born in July. Thus, finally we observe a total of 269,392 women in our sample.

We use this database as a robustness check of the impact of the reform on some fertility outcomes (see Section 3.1) . Here we define the main fertility variables used throughout the paper whose descriptive statistics could be found in Table 3.A4:

- **Probability of having a child**: Dummy variable that is equal to one if the woman had at least one children, and zero otherwise.
- **Total number of children**: Total number of children that each woman has.
- **Probability of having 3 or more children**: Dummy variable that is equal to one if the woman had at least 3 children, and zero otherwise.

Marriage Statistics

This database contains administrative data from marriage certificates for the universe of marriages held in Spain between 1976 and 2012. The information is self-reported as it comes from the Statistical Marriage Bulletins, that is filled out by the spouses at the time of registering this demographic event in the Civil Register.

The raw microdata contains 7,727,917 marriages. We, then, restrict the sample to marriages of Spanish women born between 1961 and 1971 that had 15 to 41 years old at the moment of the wedding. We also drop marriages of women born in 1966 and who therefore turned 14 the year the reform took place (1980) and those of women born in July. Thus, finally we observe a total of 2,389,673 marriages or 2,322,361 first marriages in our sample.

We use this database to assess the effect of the reform on some marriage outcomes (See Section 3.1.1). Here we define the main marriage variables used throughout the paper whose descriptive statistics could be found in Table 3.A5:

- **Age**: Age of the women when they married for the first time.
- **Number of first marriages per women in each cohort**: Total number of first marriages for a cohort of women divided by the total number of women of that same cohort multiplied by 1000.
- **Number of marriages per women in each cohort**: Total number of marriages for a cohort of women divided by the total number of women of that cohort multiplied by 1000.

Spanish National Health Survey of 2006

This database if a representative nationwide cross-sectional survey that collects health related information as well as the socio-economic status of children and adults.

The raw microdata contains 29,478 individuals. We, then, restrict the sample to Spanish women born between 1961 and 1971. We also drop women born in 1966 and who therefore turned 14 the year the reform took place (1980) and those born in July. Thus, finally we observe a total of 3,151 women in our sample.

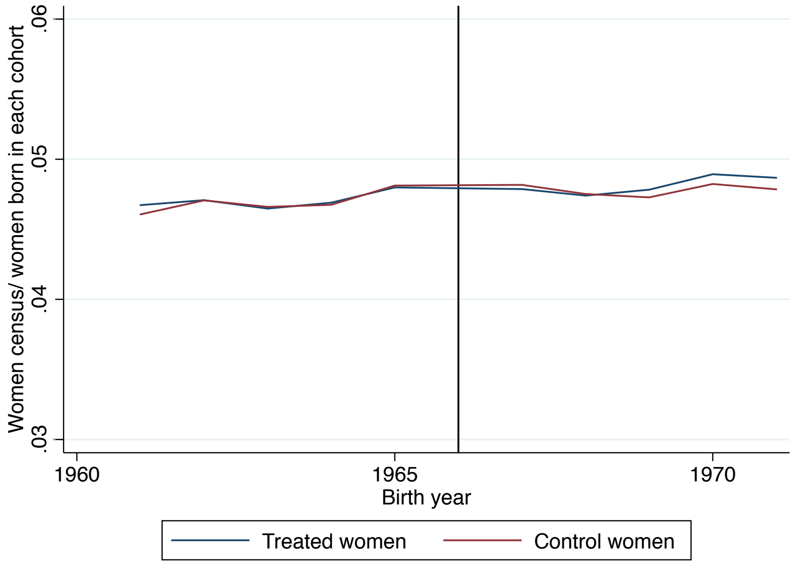
We use this database to assess the effect of the reform on some health behavior outcomes (See Section 3.2.1). Here we define the main marriage variables used throughout the paper whose descriptive statistics could be found in Table 3.A6:

- **Smoke/day**: A dummy variable that is equal to one if the woman smokes at least one cigarette a day, and zero otherwise.

- **Smoke regular**: A dummy variable that is equal to one if the woman smokes at least one cigarette a day, and zero otherwise.
- **Ex-smoker**: A dummy variable that is equal to one if the woman is an ex-smoker, and zero otherwise.
- **Pregnancy as motive for being ex-smoker**: A dummy variable that is equal to one if the woman quitted smoking during pregnancy conditional on being and ex-smoker , and zero otherwise.

Appendix Tables and Figures

Figure 3.A1: Proportion of Women Represented in the Census by Cohort and Treatment Status



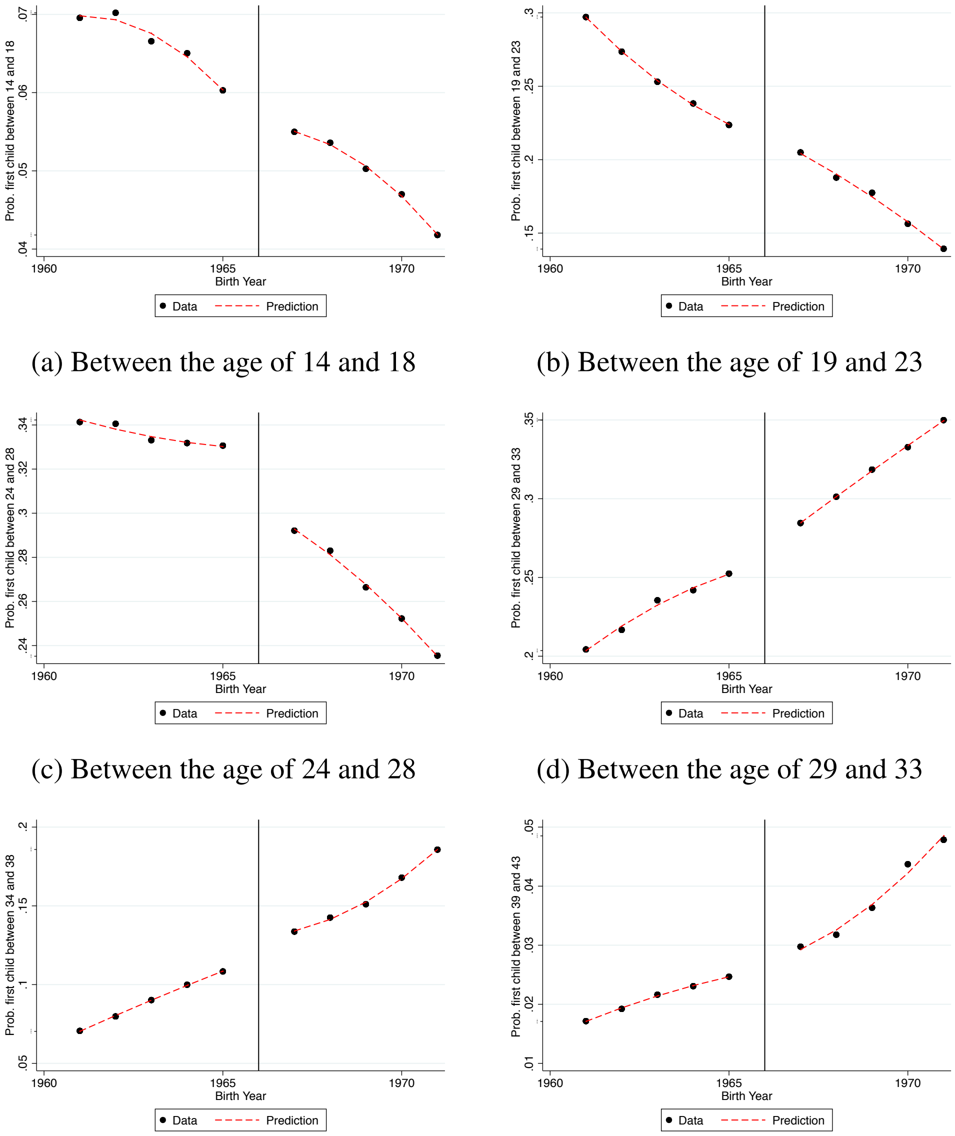
Notes: The ratio is the proportion of treated and control women within each cohort that are represented in the Census of 2001. *Treated* are women born from January to May, and *control* are those born from July to December. *Source*: Census 2001 and Birth Statistics, all women from cohorts 1961-1971.

Table 3.A1: MISSING BIRTH WEIGHT

	No birth weight (1)
Treated	-0.000 (0.000) [0.482]
Treated* Post Reform	0.001 (0.001) [0.154]
Observations	2,173,325
R^2	0.067
BirthYear FE	YES
CalendarYear FE	YES
Region FE	YES

Notes: The dependent variables is the probability that the first birth has not registered birth weight. The regression includes cohort, calendar time, and region dummies. *Treated* are individuals born from January to May, and *control* are those born from July to December. Robust standard errors clustered at the cohort level in parentheses and the p-value of the wild bootstrap with 1000 replications in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. *Source*: Birth Statistics (1980-2014), all women from cohorts 1962-1971.

Figure 3.A2: Probability of Having the First Child at a Certain Age Bracket



(a) Between the age of 14 and 18

(b) Between the age of 19 and 23

(c) Between the age of 24 and 28

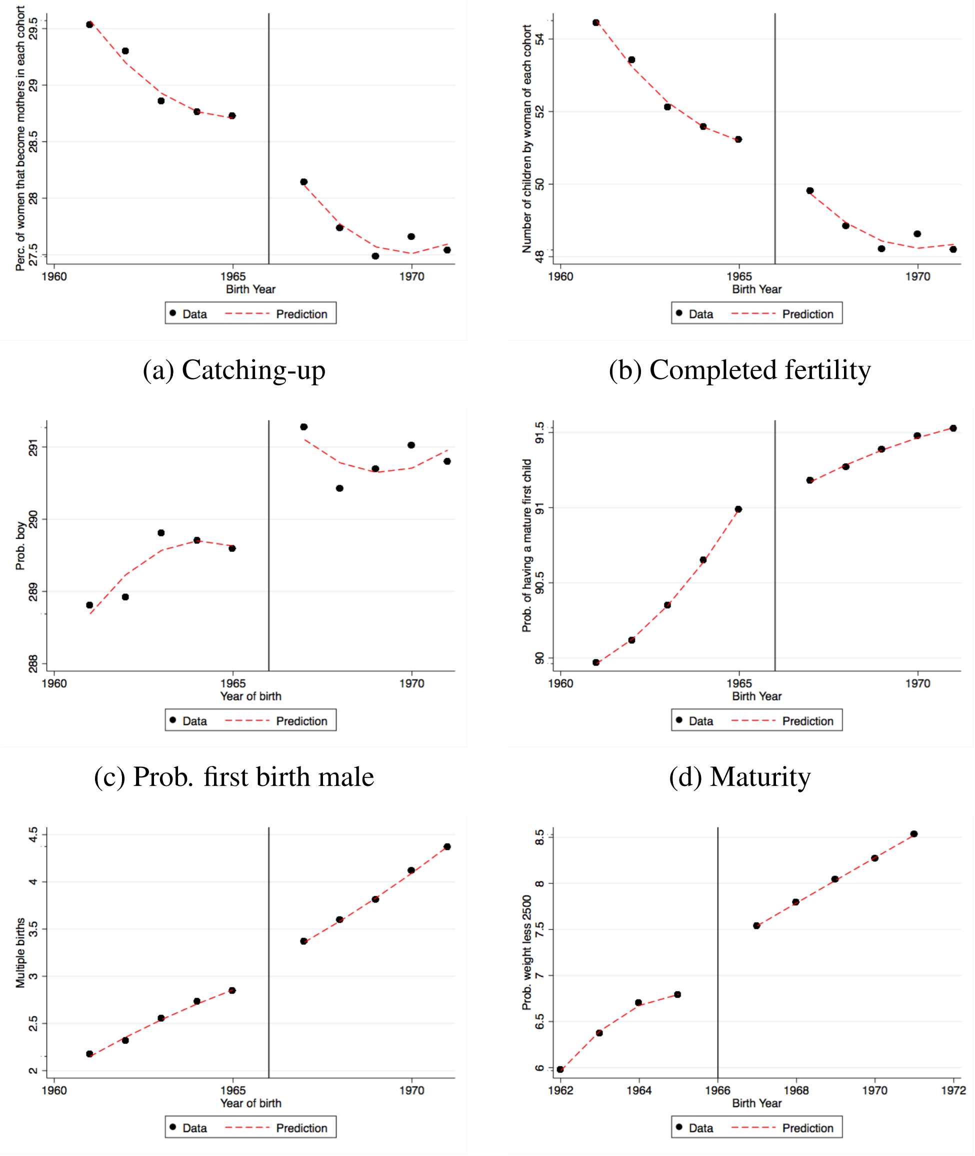
(d) Between the age of 29 and 33

(e) Between the age of 34 and 38

(f) Between the age of 39 and 43

Notes: The predictions are from a regression (with linear and quadratic trends) of the probability of women of having the first child (a) between the age of 14 and 18, (b) between the age of 19 and 23, (c) between the age of 24 and 28, (d) between the age of 29 and 33, (e) between the age of 34 and 38, and (f) between the age of 39 and 43. We consider the cohorts from 1961 to 1965 to be the cohorts before the reform and cohorts from 1966 to 1971 for after the reform. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1971 that had a child for the first time.

Figure 3.A3: Impact of the Reform on Completed Fertility and Infant Health



(a) Catching-up

(b) Completed fertility

(c) Prob. first birth male

(d) Maturity

(e) Multiple births

(f) Birth weight less 2,500 grams

Notes: The predictions are from a regression (with linear and quadratic trends) of (a) the percentage of women in each cohort that have at least one child, (b) total number of children per women of each cohort, (c) probability of having a first birth boy, (d) probability of having a first child with less than 37 gestational weeks, (e) probability of having multiple births, and (f) probability of having a first child that weight less than 2,500 grams. We consider the cohorts from 1961 to 1965 to be the cohorts before the reform and cohorts from 1966 to 1971 for after the reform. *Source*: Birth registries (1975-2014), all women from cohorts 1961-1971 that had a child for the first time.

Table 3.A2: DESCRIPTIVE STATISTICS OF THE SPANISH LABOUR FORCE SURVEY

	Treatment 1					Treatment 0				
	Observations	Mean	Std. Dev	Min.	Max.	Observations	Mean	Std. Dev	Min.	Max.
Women: Primary school or more	69924	0.97	0.17	0	1	81678	0.97	0.16	0	1
Women: Secondary school or more	69924	0.86	0.35	0	1	81678	0.87	0.34	0	1
Men: Primary school or more	72261	0.97	0.17	0	1	84282	0.97	0.16	0	1
Men: Secondary school or more	72261	0.86	0.35	0	1	84282	0.87	0.34	0	1
Women: Work	69924	60.67	48.85	0	100	81678	61.13	48.75	0	100
Women: High skill job	42425	25.86	43.79	0	100	49927	25.53	43.61	0	100
Women: Low skill job	42425	22.55	41.79	0	100	49927	21.89	41.35	0	100
Women: Part time job	26796	22.16	41.54	0	100	31553	21.92	41.37	0	100

Source: Spanish Labour Force Survey (2000-2013), for spanish women and men from cohorts 1958-1974, except the cohort of 1966.

Table 3.A3: DESCRIPTIVE STATISTICS OF THE BIRTH STATISTICS

	Treatment 1					Treatment 0				
	Observations	Mean	Std. Dev	Min.	Max.	Observations	Mean	Std. Dev	Min.	Max.
Age women when first child	1141639	27.74	5.75	14	43	1351468	27.79	5.73	14	43
First birth between 14 to 18	1141639	0.06	0.24	0	1	1351468	0.06	0.24	0	1
First birth between 19 to 23	1141639	0.22	0.41	0	1	1351468	0.21	0.41	0	1
First birth between 24 to 28	1141639	0.30	0.46	0	1	1351468	0.31	0.46	0	1
First birth between 29 and 33	1141639	0.27	0.44	0	1	1351468	0.27	0.44	0	1
First birth between 34 and 38	1141639	0.12	0.32	0	1	1351468	0.12	0.32	0	1
First birth between 39 and 43	1141639	0.03	0.17	0	1	1351468	0.03	0.17	0	1
Prob. male	1141639	29.04	29.83	0	600	1351468	29.16	29.84	0	100
Maturity	1141639	90.82	28.87	0	100	1351468	90.95	28.69	0	100
Survive 24h	1128457	99.80	4.51	0	100	1341126	99.80	4.46	0	100
Birth Weight	991899	3200.20	513.64	500	6500	1185891	3196.75	513.11	500	6400
Weight less 2500	991899	7.12	25.71	0	100	1185891	7.26	25.95	0	100
Number of first births by cohort and treatment	10	114163.90	3378.28	109450	120988.00	10	135146.80	3791.28	131197.00	142065
Number of total births by cohort and treatment	10	207696.50	9677.02	193158.00	222244.00	10	244312.10	9931.51	231494.00	258551

Source: Birth Statistics (1975-2014), all births of Spanish women from cohorts 1961-1971, except the cohort of 1966.

Table 3.A4: Descriptive Statistics of the Population and Housing Census of 2011

	Treatment 1					Treatment 0				
	Observations	Mean	Std. Dev	Min.	Max.	Observations	Mean	Std. Dev	Min.	Max.
Percentage of woman with children	136170	0.80	0.40	0	1	160290	0.80	0.40	0	1
Number of children by woman	136170	1.53	1.02	0	17	160290	1.52	1.01	0	18
Percentage of woman with more than 3 children	136170	0.12	0.32	0	1	160290	0.11	0.31	0	1

Source: Population and Housing Census (2011), for Spanish women from cohorts 1961-1971, except the cohort of 1966.

Table 3.A5: Descriptive Statistics of the Marriage Statistics

	Treatment 1					Treatment 0				
	Observations	Mean	Std. Dev	Min.	Max.	Observations	Mean	Std. Dev	Min.	Max.
Age women when first marriage	1062004	25.5	4.9	15	41	1260357	25.5	4.9	15	41
Number of first marriages by cohort and treatment	10	106200.4	3325.5	100871	112261	10	126035.7	3951.7	121942	132402
Number of total marriages by cohort and treatment	10	109216.6	3246.6	104271	115196	10	129750.7	3927.3	125034	135854

Source: Marriage Statistics (1975-2012), all marriages of Spanish women from cohorts 1961-1971, except the cohort of 1966.

Table 3.A6: Descriptive Statistics of the Spanish National Health Survey of 2006

	Treatment 1					Treatment 0				
	Observations	Mean	Std. Dev	Min.	Max.	Observations	Mean	Std. Dev	Min.	Max.
Smoke/day	1445	0.32	0.47	0	1	1706	0.32	0.47	0	1
Smoke regular	1445	0.35	0.48	0	1	1706	0.35	0.48	0	1
Ex-smoker	1445	0.22	0.42	0	1	1706	0.23	0.42	0	1
Pregnancy as motive for being ex-smoker	1445	0.05	0.22	0	1	1706	0.04	0.20	0	1

Source: Spanish National Health Survey (2006), all Spanish women from cohorts 1961-1971, except the cohort of 1966.

Bibliography

Almond, Douglas, Hilary W Hoynes, and Diane Whitmore Schanzenbach, “Inside the war on poverty: The impact of food stamps on birth outcomes,” *The Review of Economics and Statistics*, 2011, *93* (2), 387–403.

Alonso-Colmenares, María Dolores, Lara Ana, Arévalo Raquél, and Ruiz-Castillo Javier, “La Encuesta de Presupuestos Familiares 1980-81,” Departamento de Economía, Universidad Carlos II de Madrid 1999.

Angrist, Joshua and Victor Lavy, “The effects of high stakes high school achievement awards: Evidence from a randomized trial,” *The American Economic Review*, 2009, pp. 1384–1414.

Angrist, Joshua D and Stacey H Chen, “Schooling and the Vietnam-era GI Bill: Evidence from the draft lottery,” *American Economic Journal: Applied Economics*, 2011, *3* (2), 96–118.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos, “Incentives and services for college achievement: Evidence from a randomized trial,” *American Economic Journal: Applied Economics*, 2009, pp. 136–163.

_ , Philip Oreopoulos, and Tyler Williams, “When Opportunity Knocks, Who Answers?,” *Journal of Human Resources*, 2014, *49* (3).

Araujo, M Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady, “Teacher Quality and Learning Outcomes in Kindergarten,” *Quarterly Journal of Economics*, 2016, *131* (3), 1415–1453.

Azmat, Ghazala and Nagore Iriberri, "The Importance of Relative Performance Feedback Information: Evidence From a Natural Experiment using High School Students," *Journal of Public Economics*, 2010, *94* (7), 435–452.

_ and _ , "The provision of relative performance feedback: An analysis of performance and satisfaction," *Journal of Economics & Management Strategy*, 2016, *25* (1), 77–110.

Baird, Jo-Anne, "What's in a Name? Experiments with Blind Marking in A-Level Examinations," *Educational Research*, 1998, *40* (2), 191–202.

Baker, George, "Distortion and Risk in Optimal Incentive Contracts," *Journal of Human Resources*, 2002, pp. 728–751.

Baker, George P, "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 1992, pp. 598–614.

Balayla, Jacques, Laurent Azoulay, and Haim A Abenhaim, "Maternal marital status and the risk of stillbirth and infant death: a population-based cohort study on 40 million births in the United States," *Women's Health Issues*, 2011, *21* (5), 361–365.

Baltussen, Guido, G Thierry Post, Martijn J Van Den Assem, and Peter P Wakker, "Random incentive systems in a dynamic choice experiment," *Experimental Economics*, 2012, *15* (3), 418–443.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Social Preferences and the Response To Incentives: Evidence From Personnel Data," *Quarterly Journal of Economics*, 2005, pp. 917–962.

_ , _ , and _ , "Team Incentives: Evidence From a Firm Level Experiment," *Journal of the European Economic Association*, 2013, *11* (5), 1079–1114.

_ , _ , _ et al., "Incentives for Managers and Inequality among Workers: Evidence from a Firm-Level Experiment," *Quarterly Journal of Economics*, 2007, *122* (2), 729–773.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton, "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India," 2016. NBER Working Paper 22746.

Barrera-Osorio, Felipe and Dhushyanth Raju, "Teacher Performance Pay: Experimental Evidence from Pakistan," *Journal of Public Economics*, 2017, *148*, 75–91.

_ , Marianne Bertrand, Leigh L Linden, and Francisco Perez Calle, "Conditional Cash Transfers in Education: Design Features, Peer and Sibling Effects Evidence from a Randomized Experiment in Colombia," *World Bank Policy Research Working Paper Series, Vol*, 2008.

Barrow, Lisa and Cecilia E Rouse, "Financial incentives and educational investment: the impact of performance-based scholarships on student time use," Technical Report, National Bureau of Economic Research 2013.

_ , Lashawn Richburg-Hayes, Cecilia Elena Rouse, and Thomas Brock, "Paying for performance: The education impacts of a community college scholarship program for low-income adults," *Journal of Labor Economics*, 2014, *32* (3), 563–599.

Behrman, Jere R and Mark R Rosenzweig, "Does increasing women's schooling raise the schooling of the next generation?," *American Economic Review*, 2002, pp. 323–334.

_ and _ , "Returns to birthweight," *Review of Economics and Statistics*, 2004, *86* (2), 586–601.

_ , Piyali Sengupta, Petra Todd et al., "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico," *Economic Development and Cultural Change*, 2005, *54* (1), 237–75.

_ , Susan W Parker, Petra E Todd, and Kenneth I Wolpin, "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools," *Journal of Political Economy*, 2015, *123* (2), 325–364.

Bellés-Obrero, Cristina and María Lombardi, "Teacher Performance Pay and Student Learning: Evidence from a Nationwide Program in Peru," *Working Paper*, 2016.

Bennett, Trude, "Marital status and infant health outcomes," *Social science & medicine*, 1992, *35* (9), 1179–1187.

Bettinger, Eric and Robert Slonim, "Patience among children," *Journal of Public Economics*, 2007, *91* (1), 343–363.

Bettinger, Eric P, "Paying to learn: The effect of financial incentives on elementary school test scores," *Review of Economics and Statistics*, 2012, *94* (3), 686–698.

Bharadwaj, Prashant, "Impact of Changes in Marriage Law Implications for Fertility and School Enrollment," *Journal of Human Resources*, 2015, *50* (3), 614–654.

Bilal, Usama, Paula Beltrán, Esteve Fernández, Ana Navas-Acien, Francisco Bolumar, and Manuel Franco, "Gender equality and smoking: a theory-driven approach to smoking gender differences in Spain," *Tobacco control*, 2015.

Black, Sandra E, Paul J Devereux, and Kjell G Salvanes, "Staying in the Classroom and out of the maternity ward? The effect of compulsory schooling laws on teenage births*," *The Economic Journal*, 2008, *118* (530), 1025–1054.

_ , _ , and Kjell Salvanes, "From the cradle to the labor market? The effect of birth weight on adult outcomes," Technical Report, National Bureau of Economic Research 2005.

Bó, Ernesto Dal, Frederico Finan, and Martín A Rossi, "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service," *Quarterly Journal of Economics*, 2013, *128* (3), 1169–1218.

Bosdriesz, Jizzo R, Selma Mehmedovic, Margot I Witvliet, and Anton E Kunst, "Socioeconomic inequalities in smoking in low and mid income countries: positive gradients among women," *Int J Equity Health*, 2014, *13*, 14.

Botelho, Fernando, Ricardo A Madeira, and Marcos A Rangel, "Racial Discrimination in Grading: Evidence from Brazil," *American Economic Journal: Applied Economics*, 2015, *7* (4), 37–52.

Bound, John and David A Jaeger, "Do Compulsory School Attendance Laws Alone Explain the Association Between Quarter of Birth and Earnings?," *Research in Labor Economics*, 2000, *19* (4), 83–108.

Bruns, Barbara and Javier Luque, *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*, World Bank Publications, 2015.

_ , Deon Filmer, and Harry Anthony Patrinos, *Making Schools Work: New Evidence on Accountability Reforms*, World Bank Publications, 2011.

Buckles, Kasey, Melanie Guldi, and Joseph Price, "Changing the Price of Marriage Evidence from Blood Test Requirements," *Journal of Human Resources*, 2011, *46* (3), 539–567.

_ , NBER Andreas Hagemann, Ofer Malamud, and NBER Melinda Morrill, "The Effect of College Education on Mortality," 2014.

Buckles, Kasey S and Daniel M Hungerman, "Season of birth and later outcomes: Old questions, new answers," *Review of Economics and Statistics*, 2013, *95* (3), 711–724.

Burgess, Simon and Ellen Greaves, "Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities," *Journal of Labor Economics*, 2013, *31* (3), 535–576.

Byrne, Julianne, Dorothy Warburton, John M Opitz, and James F Reynolds, "Male excess among anatomically normal fetuses in spontaneous abortions," *American journal of medical genetics*, 1987, *26* (3), 605–611.

Cadena, Brian C and Benjamin J Keys, "Human capital and the lifetime costs of impatience," *American Economic Journal: Economic Policy*, 2015, *7* (3), 126–153.

Calsamiglia, Caterina and Annalisa Loviglio, "Maturity and School Outcomes in an Inflexible System: Evidence from Catalonia," 2016. Working Paper.

Camerer, Colin F and Robin M Hogarth, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 1999, *19* (1), 7–42.

Cameron, Judy and W David Pierce, "Reinforcement, reward, and intrinsic motivation: A meta-analysis," *Review of Educational research*, 1994, *64* (3), 363–423.

Carrell, Scott E and Bruce Sacerdote, *Late interventions matter too: The case of college coaching New Hampshire* number w19031, National Bureau of Economic Research Cambridge, MA, 2013.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F Halsey Rogers, "Missing in Action :Teacher and Health Worker Absence in Developing Countries," *Journal of Economic Perspectives*, 2006, *20* (1), 91–116.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 2014, *104* (9), 2633–2679.

Chevalier, Judith and Glenn Ellison, "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, 1997, *105* (6), 1167–1200.

Chong, Alberto, Isabelle Cohen, Erica Field, Eduardo Nakasone, and Maximo Torero, "Iron Deficiency and Schooling Attainment in Peru," *American Economic Journal: Applied Economics*, 2016.

Contreras, Dante and Tomás Rau, "Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile," *Economic Development and Cultural Change*, 2012, *61* (1), 219–246.

Cook, C Justin and Jason M Fletcher, "Understanding heterogeneity in the effects of birth weight on adult cognition and wages," *Journal of Health Economics*, 2015, *41*, 107–116.

Cornwell, Christopher M, Kyung Hee Lee, and David B Mustard, “Student responses to merit scholarship retention rules,” *Journal of Human Resources*, 2005, *40* (4), 895–917.

Currie, Janet and Enrico Moretti, “Mother’s Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings,” *Quarterly Journal of Economics*, 2003, pp. 1495–1532.

Cygan-Rehm, Kamila and Miriam Maeder, “The effect of education on fertility: Evidence from a compulsory schooling reform,” *Labour Economics*, 2013, *25*, 35–48.

Dahl, Gordon B, “Early teen marriage and future poverty,” *Demography*, 2010, *47* (3), 689–718.

Dearden, Lorraine, Carl Emmerson, Christine Frayne, and Costas Meghir, “Conditional cash transfers and school dropout rates,” *Journal of Human Resources*, 2009, *44* (4), 827–857.

Deaton, Angus and Nancy Cartwright, “Understanding and Misunderstanding Randomized Controlled Trials,” 2016. National Bureau of Economic Research Working Paper w22595.

Deci, Edward L, “The effects of contingent and noncontingent rewards and controls on intrinsic motivation,” *Organizational behavior and human performance*, 1972, *8* (2), 217–229.

_ and Richard M Ryan, *Intrinsic motivation*, Wiley Online Library, 1975.

Dee, Thomas S, “Are there civic returns to education?,” *Journal of Public Economics*, 2004, *88* (9), 1697–1720.

_ , “Conditional cash penalties in education: Evidence from the Learnfare experiment,” *Economics of Education Review*, 2011, *30* (5), 924–937.

Deserranno, Erika, “Financia Incentives as Signals: Experimental Evidence From the Recruitment of Health Workers,” 2016. Working Paper.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using randomization in development economics research: A toolkit," *Handbook of development economics*, 2007, *4*, 3895–3962.

Eckstein, Zvi and Kenneth I Wolpin, "Why youths drop out of high school: The impact of preferences, opportunities, and abilities," *Econometrica*, 1999, *67* (6), 1295–1339.

Edmonds, Eric V and Maheshwor Shrestha, "The Impact of Minimum Age of Employment Regulation on Child Labor and Schooling: Evidence from UNICEF MICS Countries," Technical Report, National Bureau of Economic Research 2012.

Ewijk, Reyn Van, "Same Work, Lower Grade? Student Ethnicity and Teachers' Subjective Assessments," *Economics of Education Review*, 2011, *30* (5), 1045–1058.

Figlio, David, Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth, "The Effects of Poor Neonatal Health on Children's Cognitive Development," *The American Economic Review*, 2014, *104* (12), 3921–3955.

Figlio, David N, "Testing, crime and punishment," *Journal of Public Economics*, 2006, *90* (4), 837–851.

_ and Joshua Winicki, "Food for thought: the effects of school accountability plans on school nutrition," *Journal of Public Economics*, 2005, *89* (2), 381–394.

Fletcher, Jason M, Jeremy C Green, and Matthew J Neidell, "Long term effects of childhood asthma on adult health," *Journal of health economics*, 2010, *29* (3), 377–387.

Fort, Margherita, "Just A Matter of Time: Empirical Evidence of the Causal Effect of Education on Fertility in Italy," 2007.

_ , Nicole Schneeweis, and Rudolf Winter-Ebmer, "More Schooling, More Children: Compulsory Schooling Reforms and Fertility in Europe," 2011.

Freedman, David A, "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 2008, *40* (2), 180–193.

Fryer, Roland G, "Financial Incentives and Student Achievement: Evidence from Randomized Trials*.," *Quarterly Journal of Economics*, 2011, *126* (4).

_ , "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," *Journal of Labor Economics*, 2013, *31* (2), 373–407.

Gaudino, James A, Bill Jenkins, and Roger W Rochat, "No fathers' names: a risk factor for infant mortality in the State of Georgia, USA," *Social science & medicine*, 1999, *48* (2), 253–265.

Geruso, Michael, D Clark, and H Royer, "The impact of education on family formation: Quasi-experimental evidence from the UK," Technical Report, mimeo, University of California, Santa Barbara 2014.

Glewwe, Paul, Nauman Ilias, and Michael Kremer, "Teacher Incentives," *American Economic Journal: Applied Economics*, 2010, *2* (3), 205–227.

Gneezy, Uri and Aldo Rustichini, "A Fine is a Price," *Journal of Legal Studies*, 2000, *29* (1).

_ and _ , "Pay enough or don't pay at all," *Quarterly journal of economics*, 2000, pp. 791–810.

_ and _ , "Pay Enough or Don't Pay at All," *Quarterly Journal of Economics*, 2000, pp. 791–810.

_ , Stephan Meier, and Pedro Rey-Biel, "When and Why Incentives (Don't) Work to Modify Behavior," *Journal of Economic Perspectives*, 2011, *25* (4), 191–209.

Goldin, Claudia and Lawrence F Katz, "Mass Secondary Schooling and the State The Role of State Compulsion in the High School Movement," *Understanding Long-Run Economic Growth: Geography, Institutions, and the Knowledge Economy*, 2011, p. 275.

Goodman, Sarena F and Lesley J Turner, “The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program,” *Journal of Labor Economics*, 2013, *31* (2), 409–420.

Groves, Theodore, Yongmiao Hong, John McMillan, and Barry Naughton, “Autonomy and Incentives in Chinese State Enterprises,” *Quarterly Journal of Economics*, 1994, pp. 183–209.

Guryan, Jonathan and Melissa S Kearney, “Gambling at lucky stores: Empirical evidence from state lottery sales,” *The American Economic Review*, 2008, *98* (1), 458–473.

Hanna, Rema N and Leigh L Linden, “Discrimination in Grading,” *American Economic Journal: Economic Policy*, 2012, *4* (4), 146–168.

Hanushek, Eric A and Steven G Rivkin, “Generalizations about Using Value-Added Measures of Teacher Quality,” *American Economic Review*, 2010, *100* (2), 267–271.

Hirshleifer, Sarojini, “Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance,” Technical Report, Working Paper 2015.

Hobel, Calvin J, Christine Dunkel-Schetter, Scott C Roesch, Lony C Castro, and Chander P Arora, “Maternal plasma corticotropin-releasing hormone associated with stress at 20 weeks’ gestation in pregnancies ending in preterm delivery,” *American journal of obstetrics and gynecology*, 1999, *180* (1), S257–S263.

Holmstrom, Bengt, “Moral Hazard in Teams,” *Bell Journal of Economics*, 1982, pp. 324–340.

_ and Paul Milgrom, “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design,” *Journal of Law, Economics, & Organization*, 1991, *7*, 24–52.

Hoynes, Hilary, Marianne Page, and Ann Huff Stevens, “Can targeted transfers improve birth outcomes?: Evidence from the introduction of the WIC program,” *Journal of Public Economics*, 2011, *95* (7), 813–827.

Imberman, Scott A and Michael F Lovenheim, "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System," *Review of Economics and Statistics*, 2015, *97* (2), 364–386.

Jacob, Brian A and Steven D Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics*, 2003, pp. 843–877.

Jiménez-Martín, Sergi, Judit Vall-Castello, and Elena del Rey, "The Effect of Changes in the Statutory Minimum Working Age on Educational, Labor and Health Outcomes," *IZA Discussion Papers*, 2015, (9092), http://ftp.iza.org/dp9092.pdf.

Jolly, Matthew, Neil Sebire, John Harris, Stephen Robinson, and Lesley Regan, "The risks associated with pregnancy in women aged 35 years or older," *Human reproduction*, 2000, *15* (11), 2433–2437.

Kahneman, Daniel and Amos Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica: Journal of the Econometric Society*, 1979, pp. 263–291.

Kandel, Eugene and Edward P Lazear, "Peer Pressure and Partnerships," *Journal of Political Economy*, 1992, pp. 801–817.

Kandori, Michihiro, "Social Norms and Community Enforcement," *Review of Economic Studies*, 1992, *59* (1), 63–80.

Kane, Thomas J and Cecilia Elena Rouse, "Labor-market returns to two-and four-year college," *The American Economic Review*, 1995, *85* (3), 600–614.

Kerwin, Jason T, Rebecca Thornton et al., "Making the Grade: Understanding What Works for Teaching Literacy in Rural Uganda," *Unpublished manuscript. University of Illinois, Urbana, IL*, 2015.

Kirby, Kris N, Gordon C Winston, and Mariana Santiesteban, "Impatience and grades: Delay-discount rates correlate negatively with college GPA," *Learning and individual Differences*, 2005, *15* (3), 213–222.

Kırdar, Murat G, “The Impact of Schooling on the Timing of Marriage and Fertility: Evidence from a Change in Compulsory Schooling Law,” Technical Report, Society for Economic Dynamics 2009.

Kohn, Alfie, *Punished by rewards: The trouble with gold stars, incentive plans, A’s, praise, and other bribes*, Houghton Mifflin Harcourt, 1999.

Koretz, Daniel M, “Limitations in the Use of Achievement Tests as Measures of Educators’ Productivity,” *Journal of Human Resources*, 2002, *37* (4), 752–777.

Kremer, Michael, Edward Miguel, and Rebecca Thornton, “Incentives to learn,” *The Review of Economics and Statistics*, 2009, *91* (3), 437–456.

Laibson, David, “Golden eggs and hyperbolic discounting,” *The Quarterly Journal of Economics*, 1997, pp. 443–477.

Landry, Craig E, Andreas Lange, John A List, Michael K Price, Nicholas G Rupp et al., “Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment,” *The Quarterly Journal of Economics*, 2006, *121* (2), 747–782.

Lassibille, Gérard and Ma Lucía Navarro Gómez, “How long does it take to earn a higher education degree in Spain?,” *Research in Higher Education*, 2011, *52* (1), 63–80.

Lavy, Victor, “Evaluating the Effect of Teachers’ Group Performance Incentives on Pupil Achievement,” *Journal of Political Economy*, 2002, *110* (6), 1286–1317.

_ , “Do Gender Stereotypes Reduce Girls’ or Boys’ Human Capital Outcomes? Evidence From a Natural Experiment,” *Journal of Public Economics*, 2008, *92* (10), 2083–2105.

_ , “Performance Pay and Teachers’ Effort, Productivity, and Grading Ethics,” *American Economic Review*, 2009, *99* (5), 1979–2021.

_ , “Teachers’ Pay for Performance in the Long-Run: Effects on Students’ Educational and Labor Market Outcomes in Adulthood,” 2015. NBER Working Paper 20983.

Lazear, Edward P and Sherwin Rosen, “Rank-Order Tournaments as Optimum Labor Contracts,” *Journal of Political Economy*, 1981, *89* (5), 841–864.

Le, Vi-Nhuan, “Should Students Be Paid for Achievement? A Review of the Impact of Monetary Incentives on Test Performance,” *NORC Working Paper Series*, 2015.

Lee, Ronald, Andrew Mason, Eugenia Amporfu, Chong-Bum An, Luis Rosero Bixby, Jorge Bravo, Marisa Bucheli, Qiulin Chen, Pablo Comelatto, Deidra Coy et al., “Is low fertility really a problem? Population aging, dependency, and consumption,” *Science*, 2014, *346* (6206), 229–234.

León, Alexis, “The Effect of Education on Fertility: Evidence from Compulsory Schooling Laws,” Technical Report, University of Pittsburgh, Department of Economics 2006.

Leridon, Henri, “Can assisted reproduction technology compensate for the natural decline in fertility with age? A model assessment,” *Human Reproduction*, 2004, *19* (7), 1548–1553.

Leuven, Edwin, Hessel Oosterbeek, and Bas Klaauw, “The effect of financial rewards on students’ achievement,” *Journal of the European Economic Association: evidence from a randomized experiment*, 2010, *8* (6), 1243–1265.

Levitt, Steven D, John A List, and Sally Sadoff, “The effect of performance-based incentives on educational achievement: Evidence from a randomized experiment,” Technical Report, National Bureau of Economic Research 2016.

_ , _ , Susanne Neckermann, and Sally Sadoff, “The behavioralist goes to school: Leveraging behavioral economics to improve educational performance,” *American Economic Journal: Economic Policy*, 2016, *8* (4), 183–219.

Lleras-Muney, Adriana, “Were Compulsory Attendance and Child Labor Laws Effective? An Analysis from 1915 to 1939,” *J. Law & Econ.*, 2002, *45*, 401–691.

Mackay, Judith and Amanda Amos, “Women and tobacco,” *Respirology*, 2003, *8* (2), 123–130.

March, Christoph, Antony Ziegelmeyer, and Ben Greiner, “Monetary incentives in large-scale experiments: A case study of risk aversion,” 2014.

McCrary, Justin and Heather Royer, “The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth,” *American Economic Review*, 2011, *101*, 158–195.

Mizala, Alejandra and Ben Ross Schneider, “Negotiating Education Reform: Teacher Evaluations and Incentives in Chile (1990–2010),” *Governance*, 2014, *27* (1), 87–109.

_ and Hugo Ñopo, “Measuring the Relative Pay of School Teachers in Latin America 1997–2007,” *International Journal of Educational Development*, 2016, *47*, 20–32.

Monstad, Karin, Carol Propper, and Kjell G Salvanes, “Education and fertility: Evidence from a natural experiment,” *The Scandinavian Journal of Economics*, 2008, *110* (4), 827–852.

Muralidharan, Karthik and Venkatesh Sundararaman, “The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India,” *Economic Journal*, 2010, *120* (546), F187–F203.

_ and _ , “Teacher Performance Pay: Experimental Evidence from India,” *Journal of Political Economy*, 2011, *119* (1), 39–77.

Neal, Derek, “The Design of Performance Pay in Education,” *Handbook of the Economics of Education*, 2011, *4*, 495–550.

_ and Diane Whitmore Schanzenbach, "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics*, 2010, *92* (2), 263–283.

Newstead, Stephen E and Ian Dennis, "Blind Marking and Sex Bias in Student Assessment," *Assessment and Evaluation in higher Education*, 1990, *15* (2), 132–139.

Niza, Claudia, Caroline Rudisill, and Paul Dolan, "Vouchers versus lotteries: what works best in promoting chlamydia screening? a cluster randomized controlled trial," *Applied economic perspectives and policy*, 2014, *36* (1), 109–124.

Onderstal, Sander, Arthur JHC Schram, and Adriaan R Soetevent, "Bidding to give in the field," *Journal of Public Economics*, 2013, *105*, 72–85.

Oreopoulos, Philip, "Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling," *Journal of public Economics*, 2007, *91* (11), 2213–2229.

Oyer, Paul, "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality," *Quarterly Journal of Economics*, 1998, pp. 149–185.

Pampel, Fred C, "Age and education patterns of smoking among women in high-income nations," *Social Science & Medicine*, 2003, *57* (8), 1505–1514.

Paola, Maria De, Vincenzo Scoppa, and Rosanna Nisticò, "Monetary incentives and student achievement in a depressed labor market: Results from a randomized experiment," *Journal of Human Capital*, 2012, *6* (1), 56–85.

Prendergast, Canice, "The Provision of Incentives in Firms," *Journal of Economic Literature*, 1999, *37* (1), 7–63.

Rockoff, Jonah E, "The Impact of Individual Teachers on Student Achievement: Evidence From Panel Data," *American Economic Review*, 2004, *94* (2), 247–252.

Rodriguez-Planas, Nuria, "Longer-term impacts of mentoring, educational services, and learning incentives: Evidence from a randomized trial in the United

States," *American Economic Journal: Applied Economics*, 2012, *4* (4), 121–139.

Schiaffino, Anna, Esteve Fernandez, Carme Borrell, Esteve Salto, Montse Garcia, and Josep Maria Borras, "Gender and educational differences in smoking initiation rates in Spain from 1948 to 1992," *The European Journal of Public Health*, 2003, *13* (1), 56–60.

Scott-Clayton, Judith, "On money and motivation a quasi-experimental analysis of financial incentives for college achievement," *Journal of Human Resources*, 2011, *46* (3), 614–646.

Shaienks, Danielle and Tomasz Gluszynski, *Participation in postsecondary education: Graduates, continuers and drop outs: Results from YITS cycle 4*, Statistics Canada Ottawa, 2007.

Shapiro, Doug, Afet Dundar, Jin Chen, Mary Ziskin, Eunkyoung Park, Vasti Torres, and Yi-Chen Chiang, "Completing College: A National View of Student Attainment Rates. Signature [TM] Report 4.," *National Student Clearinghouse*, 2012.

Silles, Mary A, "The effect of schooling on teenage childbearing: evidence using changes in compulsory education laws," *Journal of Population Economics*, 2011, *24* (2), 761–777.

Smith, James P, "The impact of childhood health on adult labor market outcomes," *The review of economics and statistics*, 2009, *91* (3), 478–489.

Springer, Matthew G, Laura Hamilton, Daniel F McCaffrey, Dale Ballou, Vi-Nhuan Le, Matthew Pepper, JR Lockwood, and Brian M Stecher, "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching.," *National Center on Performance Incentives*, 2010.

Steinberg, Laurence, Sandra Graham, Lia O'Brien, Jennifer Woolard, Elizabeth Cauffman, and Marie Banich, "Age differences in future orientation and delay discounting," *Child development*, 2009, *80* (1), 28–44.

Tran, Anh and Richard Zeckhauser, “Rank as an inherent incentive: Evidence from a field experiment,” *Journal of Public Economics*, 2012, *96* (9), 645–650.

Volpp, KG, LK John, AB Troxel, L Norton, J Fassbender, and G Loewenstein, “Financial incentive-based approaches for weight loss: a randomized trial.,” *JAMA*, 2008, *300* (22), 2631–2637.

Wolpin, Kenneth I, “Determinants and consequences of the mortality and health of infants and children,” *Handbook of Population and Family Economics*, 1993, *1*, 483–557.

y Deporte Ministerio de Educación, Cultura, “Datos básicos del sistema universitario español. Curso 2013-2014.,” Technical Report, Ministerio de Educación, Cultura y Deporte 2015.

Zimmerman, Seth D, “The returns to college admission for academically marginal students,” *Journal of Labor Economics*, 2014, *32* (4), 711–754.