Discussion Paper Series – CRC TR 224

Discussion Paper No. 126
Project A 02

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

Cristina Bellés-Obrero*
María Lombardi**

September 2019

*University of Mannheim, Department of Economics, Office 326, L7, 3-5, 68161, Mannheim, Germany & CRES (Universitat Pompeu Fabra). *Email:* cbelleso@mail.uni-mannheim.de
**Universidad Torcuato Di Tella, School of Government, Av. Figueroa Alcorta 7350, C1428BCW Buenos Aires, Argentina. *Email:* mlombardi@utdt.edu

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

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

Cristina Bellés-Obrero†
University of Mannheim

María Lombardi‡
Universidad Torcuato Di Tella

September 2019

Abstract

We study the impact on student achievement of a nationwide teacher pay-for-performance program implemented in Peruvian public secondary schools in 2015. Schools compete in a tournament primarily based on 8th graders' performance in a standardized test, where the principal and teachers of the top 20 percent of schools receive a substantial bonus. We perform a difference-in-differences estimation comparing the internal grades of 8th and 9th graders of the same school, before and after the program. We find a precisely estimated zero effect on student achievement, and we reject impacts greater than 0.017 standard deviations, well below those previously found in the literature. We provide evidence against a series of potential explanations, and argue that this zero effect could be a consequence of teachers' uncertainty about how to improve their students' performance in the standardized test tied to the bonus.

JEL Classifications: I21, M52, J4

Keywords: education, teachers, incentives, compensation, Peru.

*We are especially grateful to Libertad González, Gianmarco León, Stephan Litschig and Alessandro Tarozzi. We also thank Ana Rute Cardoso, Ana Figueiredo, Niklas Heusch, Victor Lavy, Ricardo Maertens, Ana Tur-Prats, Tomás Williams, and conference and seminar participants for their helpful comments. We are also grateful to the Peruvian Ministry of Education for providing and assisting with the data used in the study. The views expressed herein are those of the authors and do not necessarily reflect those of the Peruvian Government. Bellés-Obrero acknowledges support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (Project A02), and the project ECO2017-82350-R.

†University of Mannheim, Department of Economics, Office 326, L7, 3-5, 68161, Mannheim, Germany & CRES (Universitat Pompeu Fabra). *Email:* cbelleso@mail.uni-mannheim.de

‡Universidad Torcuato Di Tella, School of Government, Av. Figueroa Alcorta 7350, C1428BCW Buenos Aires, Argentina. *Email:* mlombardi@utdt.edu

1 Introduction

Teacher quality is one of the key factors determining student achievement. Individuals exposed to better teachers perform better in school (Rockoff, 2004; Hanushek and Rivkin, 2010; Araujo et al., 2016), and are 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 may discourage high-skilled individuals from entering 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, effort, and ultimately student learning, academics and policymakers have proposed linking teachers' pay to their students' performance. The impact of teacher pay-for-performance programs has been examined in low- and middle-income countries such as India, Pakistan, Kenya, China, Chile, and Mexico. Yet the majority of these programs were run by NGOs at a relatively small scale, and so their conclusions do not necessarily apply to nationwide, government-led interventions.

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 public secondary schools compete within a group of comparable schools on the basis of their annual performance. Schools are deemed comparable if they belong to same school district, have the same school-day length, and are also urban or rural. The principal and teachers in schools that are ranked in the top 20% within their BE group obtain a fixed payment, amounting to over one month's salary. The incentives provided by BE are collective (at the school-level), since all teachers are rewarded if their school wins. However, the main performance measure used to rank schools is their average score in math and language standardized tests, taken by 8$^{\text{th}}$ graders only. This feature of the program, which we exploit in our identification, implies that a school's probability of obtaining the bonus in 2015 hinges on the achievement of 8$^{\text{th}}$ grade students. Our estimation relies on a novel administrative database collected by the Peruvian Ministry of Education, which covers all students in 2013-2015 and contains annual information on the grades that students receive from their teachers in every subject (their "internal grades"). The availability of achievement measures for students in all grades allows us to compare the change in internal grades of 8^{th} graders with that 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. Since internal grades do not have a direct impact on a school's BE score, teachers' grading tactics should not be influenced by the incentive. 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.

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). Separate estimations in each of the 409 groups in which schools compete reveal zero average effects in the majority of cases, providing additional evidence of the null effect of BE on student achievement. Due to the tournament nature of BE, it is possible that only teachers from schools close to the margin are incentivized, 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. We also assess whether BE improved student achievement in our comparison group by using data on the overlap of teachers in 8^{th} and 9^{th} grade, and discard the existence of any spillovers that could bias our estimates.

Why was student learning unaffected by the BE program?[1] We provide suggestive evidence that the null effect is not a result of the size or collective nature of the incentive, nor is it driven by teachers being uninformed about the program or only focusing on increasing standardized test scores – the incentivized outcome – without influencing student learning in a meaningful way.

[1]To answer this question, we carried out an online survey on a sample of public secondary school teachers. However, given the potential bias in teachers' responses, we take the results from this survey with caution.

We argue that 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 are not tested at baseline and these tests were performed for the first time in 2015, teachers may not have known what pedagogical practices lead to higher test scores. And because students have no stakes in these evaluations, the mapping between between teachers' effort and their chances of winning the bonus may be too weak.

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.[2] Using a randomized controlled trial in rural schools in the Indian state of Andhra Pradesh, Muralidharan and Sundararaman (2011) find that providing individual and collective monetary incentives to teachers based on their students' test score gains had a significant and sizable effect on 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 raised students' test scores in the exams tied to the incentive, it had no impact on non-incentivized exams covering similar topics.[3] Behrman et al. (2015) ran a randomized controlled trial in 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 although incentivizing teachers had no impact on students' test scores, there was a significant increase in student performance when students themselves were incentivized, and even greater effects when both students and teachers were

[2] In the context of a high-income country, the quasi-experimental studies of Lavy (2002) and Lavy (2009) in Israeli high schools find positive and significant impacts on different measures of student performance tied to the incentives. Fryer (2013), and Goodman and Turner (2013) analyze a collective teacher incentive program in approximately 200 New York City public schools, and find no evidence of increased student attainment or changes in student or teacher behavior. Finally, Springer et al. (2010) conducted a three-year study in Nashville, and find a positive effect on standardized test scores only among teachers instructing the same set of students in multiple subjects.

[3] Although direct observation of teachers in Muralidharan and Sundararaman (2011) shows no impact 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) 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 test-preparation classes, suggesting that teachers' efforts might have narrowly targeted the incentivized outcome.

given incentives.[4] Barrera-Osorio and Raju (2017) evaluate a pay-for-performance program in low-performing primary schools in Pakistan, and find no impact on student scores in government exams. The authors argue that the high volatility in government test scores may have dampened teachers' incentives to increase their effort (Barrera-Osorio and Ganimian, 2016). The closest paper to ours is 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 received higher scores in math and language standardized tests, as compared to students in private schools that were not eligible for the bonus.

One contribution of our paper is that we examine whether teacher incentives can work in the context of a scaled-up, national intervention. Except for Contreras and Rau (2012), all other studies in the literature focus on smaller interventions. Barrera-Osorio and Raju (2017) evaluate a pay-for-performance program implemented in 450 schools, 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 reached more than 8,000 secondary schools across Peru, providing incentives to approximately 86,000 teachers responsible for instructing 8$^{\text{th}}$ grade students in these schools. Perhaps more importantly, with the exception of Barrera-Osorio and Raju (2017), the program design and implementation in the aforementioned studies was carried out by an NGO, whereas BE was directly implemented by the Peruvian Ministry of Education. While these experimental studies provide proof-of-concept of whether teacher pay-for-performance can increase student achievement, they face external validity issues, as in any randomized controlled trial (Deaton and Cartwright, 2018), making their findings not necessarily generalizable to a large-scale program. This notion is underscored by Banerjee et al. (2016) and Bold et al. (2018), who show that successful educational interventions led by NGOs do not yield the same impact when scaled-up and implemented within the existing educational system. Budgetary constraints (Kerwin and Thornton, 2015) and opposition from teacher unions (Mizala and Schneider, 2014) make several aspects

[4]In line with these findings, incentivized students reported exerting higher effort in preparing for the exam. The self-reported behavior of teachers is not as consistent with the paper's findings, since teachers in all three treatment arms were more likely to report having prepared their students for the test, both in and out of class.

of these types of interventions unfeasible in a nationwide program. Take for instance the most successful intervention in this literature, the Muralidharan and Sundararaman (2011) study. In this program, students were evaluated at the start of the school year, their teachers received detailed feedback on the performance of each student and suggestions on how to use this feedback to improve learning, and NGO workers made several visits to the schools to monitor the attendance and classroom practices of teachers. Frequently testing students, providing such detailed feedback, and regularly monitoring teachers may be too costly to implement at scale by the government of a developing country. It is thus crucial for policymakers to better understand what role these and other features of teacher pay-for-performance programs play in their success. While we cannot fully conclude which characteristics of the BE contributed to its null impact, we provide some suggestive evidence, hopefully shedding more light on this discussion.

Another novelty of our study is that we use a measure of student achievement that captures the skills of students that the program targets, but is not 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.[5] An identification strategy relying on standardized test scores (or other incentivized indicators) as an outcome cannot fully disentangle whether an improvement in student performance is the consequence of higher learning or the result 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 students performed better in the tests used to award the bonus, but find no effect in their performance in an alternative exam not linked to the incentive. With the exception of the latter, all other papers in the literature assess student achievement using measures of learning which are directly targeted by the incentive.[6] Although it would

[5]Internal grades are completely independent of standardized test scores in Peru. Since students' individual scores are never reported, it is impossible for teachers to give any weight to these tests in their internal assessment. And even though teachers could potentially use internal grading as a tool to motivate students to do well in the standardized tests, internal grades are assigned after the tests are administered.

[6]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 a standardized test to include mechanical and conceptual questions; while performance in the former can be easily affected by a teacher coaching his students

also be interesting to analyze the impact of BE on students' scores in the standardized test tied to the bonus, there is no comparison group to perform a causal analysis. Using internal grades as our outcome has the advantage of capturing students' performance without directly influencing teachers' probability of obtaining the bonus. Even though this measure may 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 describes the educational system in Peru and the Peruvian teacher pay-for-performance program. Section 3 discusses our strategy for estimating the effect of teacher incentives on student performance, and Section 4 describes the data. Section 5 presents our main results, and Section 6 provides evidence on the validity of our identification strategy. Section 7 discusses the potential reasons for our findings, and Section 8 concludes.

2 Secondary Schooling in Peru and the BE Program

2.1 Secondary Schooling in Peru

Compulsory schooling in Peru lasts 12 years, and is composed of preschool, primary and secondary schooling, lasting one, six, and five years, respectively. 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, reaching 18% of all public secondary schools in 2016. While a significant portion of the student body attends private secondary schools, public institutions dominate by far. In 2014, for instance, 64% of high schools were publicly run, instructing 74% 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 high portion of students attending high school do not possess the minimum required levels of knowledge. In the 2012 round of the OECD's Programme for International Student Assessment

for the test, conceptual questions are harder to influence using these types of tactics.

(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. In 2015, salaries for the former are divided into eight pay scales, with monthly salaries ranging from 1451 soles ($\approx$ 439 dollars) to 3773 soles ($\approx$ 1142 dollars).[7] The monthly salary of contract teachers on the other hand was fixed at 1244 soles ($\approx$ 370 dollars).[8] 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, approximately 444 dollars.[9] The working week for public secondary school teachers in 2014 consisted of 26 hours, 24 of which were to be spent teaching.[10] However, as reported in a nationally representative survey in 2014, teachers spent an average of 12 hours a week performing other work-related 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.

As compared to Peruvian workers with similar qualifications and teachers in other Latin American countries, Peruvian teachers are poorly paid. A 2010 study on the salaries of teachers in Latin America finds that adjusting for the number of hours worked, Peruvian teachers earned 10% less than other Peruvian professional workers with similar levels of education (Bruns and Luque, 2015).[11] Even though Peruvian teachers are poorly paid, absenteeism is relatively low compared

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

[8]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.

[9]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*).

[10]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.

[11]The gap in teachers' pay was worse in Peru than in Mexico, Honduras, El Salvador, Costa Rica, Uruguay and Chile, but relatively better than than in Panama, Brazil and Nicaragua. Mizala and Ñopo (2016) examine the patterns of teacher pay in several Latin American countries 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.

to other developing countries. Around the time at which BE was implemented, average teacher absenteeism in public schools was below 7%.[12]

2.2 The BE Program

In 2013, the Peruvian Ministry of Education launched *Bono Escuela* (BE), an annual and nation-wide teacher pay-for-performance program in public schools. The program was first implemented in primary schools, and then extended to secondary schools in 2015. Secondary schools, the focus of this paper, were only included in BE as of 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. 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.[13] In secondary schools, only 8th graders are tested. The ECE is implemented by the Peruvian National Statistics Institute, 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.

BE is a tournament providing collective incentives, such that the principal and every teacher in a school are rewarded if their school wins.[14] To ensure that schools competing against each other are comparable, they are separated into groups by school district, school day length, and by whether they are urban or rural. There are 409 groups in total, with an average of 36 schools per group.

[12] These figures were obtained from the first round of *Semaforo Escuela*, a program launched around April 2015 in which trained enumerators make periodic visits to public schools, and register information on teacher, student, and director absenteeism. 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).

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

[14] Most of the interventions in this literature provide collective incentives. Examples include Lavy (2002) in Israel, Glewwe et al. (2010) in Kenya, one of the treatment arms in Muralidharan and Sundararaman (2011) in India, Contreras and Rau (2012) in Chile, and Fryer (2013) and Goodman and Turner (2013) in the US.

Teachers in schools in the top 10% of their group receive a bonus of 2000 soles (approximately 605 dollars), whereas those in the top 10%-20% are paid 1500 soles (454 dollars). Every teacher in a winning school receives the same bonus, but the school principal receives a slightly larger payment (500 extra soles). The bonus is paid out in November of the following year, together with the salary of that month. 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.[15]

As outlined in Table 1, a school's BE score in 2015 was composed of several factors. The school's average grade in the ECE standardized tests taken by 8th graders makes up for 40% of the score. In order to prevent teachers from encouraging the absenteeism of low achieving students on the day of the assessment, schools not complying with a minimum rate of student participation are disqualified from taking part in BE. The participation rate in the ECE must be 80%, 90% and 95% in schools with only one, two, or more than two 8th grade classes, respectively.[16] Additionally, 35% of weight is given to the entire school's intra-annual retention rate, calculated as the proportion of enrolled students still in school at the end of the year, without considering those who transferred to another institution. In 2014, the year before BE was implemented, public secondary schools had an average retention rate over 99%, and less than 6% had retention rates below 95%. Intra-anual retention rates are so high because most of the dropouts occur after the school year ends.[17] As schools had very little leeway for increasing their rates of retention, they could not compete on the basis of this indicator.[18] We believe that the Ministry of Education gave such a large weight to this indicator to counteract any perverse incentives schools may have to improve their test scores by encouraging their weakest students to drop out (Gilligan et al., 2018). An extra 5% of the

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

[16]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.

[17]Approximately 6% of the students enrolled in 7th to 10th grade of a public school in 2014 dropped out over the summer. Furthermore, more than 65% of the students who in 2014 failed 11th grade, the last year of school, did not re-enroll in 2015.

[18]Average retention rates only improved by 0.29 percentage points between 2014 and 2015.

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. The remaining 20% of the score depends on an index of school management, composed of measures of teacher attendance, management of school infrastructure, compliance with class hours, pedagogical practices, and learning environment. The first three indicators are measured at the school-level, and are collected by independent evaluators making visits to all public schools, whereas the last two apply to 8^{th} grade only, and are obtained from questionnaires handed out to students during the ECE. On the whole, around 80% of a school's score ultimately depends on the actions of 8^{th} grade math and language teachers.[19] 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.

The timing of BE is depicted in Figure 1. The school year in Peru starts in March and ends in December. At the end of 2014, once the implementation of the ECE test in secondary schools was confirmed for 2015, the Minister of Education announced the possibility of extending the BE program to secondary schools.[20] The government resolution regulating the 2015 BE came out on July 25^{th}, almost four months before 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.

3 Estimation Strategy

We exploit the fact that a school's score for the BE primarily depends on the performance of 8^{th} grade students in the math and language ECE tests for estimating the causal effect of BE on student

[19]The portion of the BE score that depends on retention rates is just a constant, since schools had no margin to improve this indicator. In fact, 80% of schools had a retention rate of 100% in 2014. Since 8^{th} grade math and language teachers could impact the BE score through the ECE test scores of their students (40% of weight) and their pedagogical practices and learning environment (12% of weight), we consider that their practices can influence up to 80% of the school's score (0.52/(1-0.35)).

[20]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).

learning. As schools have a much higher incentive to improve the learning of 8^{th} grade students as compared to students from other grades, 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.[21] In our preferred specification, we use a repeated cross-section of 8^{th} and 9^{th} grade students in public secondary schools, and run the following regression:

$$\textit{Internal Grade}_{ist} = \beta_0 + \beta_1\, 8^{\text{th}}\, Grade_{ist} + \beta_2\, 8^{\text{th}}\, Grade_{ist} \times Post_t + X_{ist}\,\delta + \gamma_t + \gamma_s + U_{ist}\ ,$$

where $Internal\ Grade_{ist}$ is the grade that student *i* in school *s* received from his/her teacher at the end of year *t*. We run separate regressions for math and language internal grades, and also estimate this equation for the average internal grade in all subjects not evaluated in the ECE, in order to examine whether BE impacted students' performance in other courses. $8^{\text{th}}\ Grade_{ist}$ is a dummy for whether student *i* from school *s* is an 8^{th} grader in year *t*, $Post_t$ takes a 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 BE was introduced. Including school fixed effects allow us to restrict our comparison to students facing the same educational environment, but who differ in their exposure to the BE.[22] 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

[21] Since students in the 7^{th} grade will be taking the ECE standardized test in 2016, thus affecting the school's score in the following edition of BE, the program could have an impact on these students. We thus use 9^{th} graders as our comparison group.

[22] Given that 8^{th} grade students in private schools take the ECE but these institutions are not eligible for the BE, we could run a differences-in-differences regression using 8^{th} graders from private schools as our comparison group, similar to what Contreras and Rau (2012) do in the case of Chile. However, as shown in Appendix Table A.1, public school students were already improving relatively faster than their private school counterparts in the year prior to the BE (i.e., the *Public* $\times$ *2014* coefficient is positively statistically significant). Perhaps more importantly, private schools can not serve as a control group because many other things changed across the public-private spectrum in 2015, such as differential investments in infrastructure and salary raises.

a few. We cluster our standard errors at the school level to allow for the within-school correlation of errors, and 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.

Unlike other studies on teacher pay-for-performance, our outcome variable 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 scores. Given that teachers' pay under BE is tied to performance in the ECE, an estimation relying on standardized test scores as an outcome would be unable to determine whether improvements are due to increased student learning or the result of short-term strategies fostering high test scores, such as targeting topics likely to be tested, coaching students on test taking strategies or cheating.[23] Since the BE bonus is tied to ECE test results, and not internal grades, teachers' stakes in the latter are not directly modified by the incentive program. Although teachers could potentially use internal grades as a tool to motivate students to do well in the ECE, internal grades are assigned *after* the tests are administered.[24] We also 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 6.3. Internal grades therefore have the advantage of capturing the skills of students targeted by the program (i.e., their basic competencies in math and language), without directly influencing teachers' probability of obtaining the bonus. It would still be interesting to study the impact of the BE on students' ECE test scores, because comparing both results would allow us to learn more about mechanisms. Unfortunately, we cannot do so because the ECE was implemented in secondary schools for the first time in 2015, and since BE covers all public secondary schools, there

[23]These behaviors are consistent with models of multi-tasking (Holmstrom and Milgrom, 1991; Baker, 1992, 2002), and have 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.

[24]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 Appendix Table A.2, teachers were equally likely to report that they decreased the difficulty of their classes when teaching students from 8$^{\text{th}}$ grade, as compared to students from other grades, and marginally more likely to report that they increased the difficulty of their classes. It is thus unlikely that the estimated impacts are driven by differential changes in the difficulty of the class.

is no appropriate comparison group.

Our main identifying assumption is that in the absence of BE, the performance of 8^{th} and 9^{th} grade students attending the same school would have evolved in the same way between 2014 and 2015. An inspection of the raw means in Figure 2 shows that this is a plausible assumption, as the grades of 8^{th} and 9^{th} grade students were on parallel trends before the implementation of BE in both math and language. We test this formally in Section 6.1.

Identifying the causal effect of BE 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. One must bear in mind, however, that since 84% of 8^{th} grade teachers also instruct 9^{th} grade, higher effort while teaching 8^{th} graders could potentially spill over to students in our comparison group and bias our estimates downwards. We show that our results are not driven by such positive spillovers by exploring the impact of the teacher incentive in schools with a low overlap between 8^{th} and 9^{th} grade teachers in Section 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 may lead schools to redirect their resources towards these grades, negatively impacting the internal grades of students in our comparison group. We show in Section 6.2 that this issue is not a concern in our setting.

4 Data and Descriptive Statistics

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 almost universal, reaching approximately 99.7% of public schools. Schools are required to enroll their students into the SIAGIE system at the start of the school year, and input the final grades of their entire student body at the conclusion of the academic year. Grading is done on a 0-20 scale, and students need to obtain at least 11 to pass a given subject. In addition to students' grades, this database holds information on students' age, gender, native tongue, parents' education, whether they live with their parents, and identifiers that 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. Of the 8,914 public secondary schools in Peru in 2015, 8,330 were eligible to participate in the ECE, as schools must have at least five 8th grade students in order to be eligible to take the test. Our SIAGIE database covers 8,297 of these schools. We then restrict our sample to the 7,878 schools that have at least one 8th and 9th grade classroom in 2013, 2014 and 2015.

Table 2 presents some characteristics of the 8th and 9th grade students attending public secondary schools in 2013-2015. The average final grade in math is 12.29 and 12.33 (out of 20) in 8th and 9th grade, and 84% and 85% of students pass this course, respectively. Students perform slightly better in language, where the average grade in 8th and 9th grade is 12.69 and 12.76, and 89% and 90% of students pass the course, respectively. 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 Peruvian, and Spanish is the first language for 84-85% of them. Only 5% and 4% of 8th and 9th graders were retained in the same grade the year before. Although causal identification does not require that 8th and 9th grade students are balanced in terms of observables, they appear to be very similar. Table 2 also shows some characteristics of the 7,878 public secondary schools in our sample. Less than half (39%) of the schools are located in rural areas. Each school has, on average, two classes per grade, and there are around 20 students per teacher in the average class. We also observe that each school has, on average, approximately 11 teachers teaching 8th and 9th grade, with around 83% of teachers in 8th (9th) grade also teaching 9th (8th) grade. Teachers are almost 42 years old on average, and 60% of them are male.

We complement our main empirical analysis with the results of an online survey conducted in conjunction with the Ministry of Education in October 2016. We collected information on the grades and subjects the teachers taught in 2005, their knowledge about the BE and its rules at that time, and their opinion about the size of the bonus. The survey 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 elicit teachers' perceptions of their school's ranking and probability of winning, and we asked teachers for their opinions of students' motivation in the standardized test tied to the BE. We received a response from 3,406 teachers (2.8% of all public secondary school teachers). However, given the potential bias in teachers' responses and the selected sample, the results from this survey must be taken with caution. We provide further information on the survey and its results in Appendix B

5 Results

Despite the fact that the monetary incentives provided by BE were sizable, we find that the program had no effect on 8^{th} graders' math and language internal grades, as shown in columns (1) and (4) of Table 3.[25] 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.[26] Our coefficients are precisely estimated zeros, allowing us to reject positive effects larger than 0.010 SD in math, and 0.017 in language, well below the positive 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, and 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 the scores of tests linked to the incentive, the authors found no impact on the outcome

[25]These regressions do not include students who dropped out during the school year and thus do not have a final grade for math or language. Importantly, dropout rates during the school year are very small (0.4%), and there is no differential impact across grades.

[26]The sample is slightly smaller when we include individual controls because these characteristics are missing for a small group of students (less than 0.5%). Since parents' education is missing for 12% of students, we do not control for this in our regressions. However, attrition is not differential across grades, and our results are robust to the inclusion of these controls.

of non-incentivized evaluations, consistent with our findings.

Since there are 409 distinct groups in which schools compete for the BE bonus, i.e., 409 different tournaments, we also evaluate the average effect of BE in every competition. Figure 3 displays the 8^{th} *Grade* x *Post* coefficients (and their 95% confidence interval) for math and language in each of these 409 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 5% of the BE groups,[27] providing further evidence of 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 may 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, 8^{th} grade teachers in all subjects, not just math and language, may exert more effort knowing that their school's score rests largely on the performance of these students.[28,29] We do find a positive and statistically significant effect of 0.011 SD on grades in non-incentivized courses, as shown in Table 4. Appendix Table A.3 breaks the results down by each of the nine non-incentivized courses; we observe a positive and statistically significant impact of 0.017 SD in social studies and 0.019 SD in human relations. Although statistically different from zero, the observed effect is very small, and well below the spillover effects found in other papers.[30] Furthermore, these results should be taken with caution because, as further discussed in Section 6.1, there is a small divergence in the trend in non-incentivized courses the year before the program was implemented.

[27]Furthermore, in only six out of the 409 BE groups, this holds simultaneously for math and language.

[28]Note that unlike studies carried out in primary school, math and language teachers are not generally responsible for teaching other subjects in secondary school.

[29]Additionally, due to complementarities, if learning would have been higher in math and language, student achievement in incentivized subjects could have increased indirectly (Muralidharan and Sundararaman, 2011).

[30]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, respectively, after only one year, an effect 10 to 13 times larger than the one we find.

5.1 Heterogeneous Effects

Tournament theory predicts that if teachers are risk neutral, have symmetric information, and if students in all schools have the same ability, all teachers should exert the same effort as a result of the incentive, and thus who is awarded the bonus should be random (Lazear and Rosen, 1981). However, if schools differ in their probability of winning, the incentive may not have the same power across the board. For example, teachers in schools in which students' pre-program levels of achievement are far from those in the top 20% could be discouraged from exerting extra effort, and schools which are almost guaranteed to win may not respond to the incentive either. This concern is partly attenuated in our setting by the fact that schools are grouped according to characteristics which are likely correlated with their students' performance, such as the number of hours of instruction students receive, whether they are urban or rural, and their school district. Nevertheless, there could still be differences between schools in the same group. Consistent with this notion, Contreras and Rau (2012) find that a teacher incentive program awarding a bonus to schools in the top 25% within their group only had a positive impact on schools above the 65^{th} percentile in the distribution of pre-program scores. Since ECE was implemented in secondary schools for the first time in 2015, we cannot accurately determine a school's pre-tournament probability of winning. As an alternative, 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. This approach is reasonable considering the findings of Mizala et al. (2007) that rankings based on average school test score levels in Chile are essentially equivalent to rankings based on schools' average SES. 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.[31] We then rank schools within their BE group according to this measure, and fully interact our baseline regression with 20 dummies indicating the percentile that each school belongs to within their BE group. Further confirming the null effect of BE, Figure 4 shows that the interaction coefficients are very small

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

in both math and language, and most of them are not statistically significant.

A second aspect to consider is that the strength of the incentive may be decreasing in the number of incentivized (i.e., 8^{th} grade math and language) teachers, since the marginal impact of an individual teacher's effort on his/her school's score decreases when there are more teachers reached by the incentive, as does the teachers' ability to monitor one another.[32] We use the number of 8^{th} grade classes and the enrollment in 8^{th} grade in 2015 as a proxy, since we do not know how many incentivized teachers each school has (our teacher database does not have information on the subject that each teacher is responsible for). 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 5. Finally, we do not find any effects by whether the school is 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 enrollment and urbanicity are not randomly assigned, and could be a proxy for something else. Although it would also be interesting to evaluate whether there are heterogeneous effects across teacher characteristics, we do not know which of the teachers assigned to each grade and class teaches math and language.

Following other papers in the literature, we also test for heterogeneous effects across gender, by whether the students' first language is Spanish, and by their parents' educational attainment. Parents' education and students' native tongue are proxies for socioeconomic status in Peru. As displayed in columns 1-3 of Table 5, we do not find any heterogeneous impacts across these dimensions. Considering that teachers may focus on certain students, and student responsiveness may vary according to prior achievement, we also test for differential impacts by 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).[33] 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

[32]Confirming this point, 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.

[33]Lagged grades are only available for students in 2014 and 2015. Importantly, if we restrict our sample to this period, results on average treatment effects do not change.

Schanzenbach, 2010). On the contrary, if obtaining the bonus depends on average scores, such as in BE, teachers will find it optimal to the target students most likely to respond to any increase in 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). Confirming the overall null impact of BE, we do not find any heterogeneity according to the prior performance of students, as shown in columns 4 and 5.[34]

6 Testing the Validity of the Identification Strategy

6.1 Parallel Trends

To test whether there is a divergence in the trends of 8th and 9th grade students in 2014, we add an interaction between the 8th 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 6. In the case of non-incentivized courses, however, there is a relative increase in 8th graders' internal grades in 2014. Although the magnitude of this change is very small (0.011 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.

6.2 No Spillovers to 9^{th} Grade Students

A crucial condition for identifying the causal impact of BE is that the performance of 9th grade students, our comparison group, is unaffected by the program. One concern is that schools might assign their best teachers to the 8th grade in 2015 to improve their chances of obtaining the bonus, biasing our estimates upward. Given our findings of a null effect of BE on student achievement, this

[34] 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 Appendix Table A.4.

is less of a concern. And since the program was only announced after the school year had started, as illustrated in Figure 1, it would have been difficult for schools to adjust teaching schedules at that point. Nonetheless, we perform some tests to learn whether teacher characteristics changed differently across grades in the year in which BE was introduced. Since we cannot identify which teachers are responsible for instructing math and language, we test whether the average characteristics of all teachers changed differentially across the 8^{th} and 9^{th} grades in 2015. As shown in Table 7, we do not find any differential changes in teacher composition across 8^{th} and 9^{th} grade classes in 2015. In particular, there is no differential change in the average age or gender, the average number of classes or schools taught by teachers, or the proportion of teachers who were teaching in the school or grade for the first time.[35]

A bigger concern given our null impacts is that BE improved teacher behavior overall, instead of only impacting the teaching of 8^{th} graders. As we explain in Section 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 84% 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 instruct 9^{th} graders (columns 1 and 4 of Table 8). Even though the possibility of spillovers depends on whether 8^{th} and 9^{th} graders have the same math and language teachers, we do not have information on the subjects taught by each teacher. We thus proxy the overlap of math and language teachers using the average overlap of all teachers across the 8^{th} and 9^{th} grade.

[35]Even though observable teacher characteristics are usually poor predictors of teacher quality, several studies find that the first years of teaching experience are important for student learning (Rivkin et al., 2005; Staiger and Rockoff, 2010; Araujo et al., 2016).

6.3 Internal Grades Reflect Learning

Unlike other studies on teacher pay-for-performance, we measure learning using students' internal grades instead of their standardized test scores.[36] As discussed in Section 3, the advantage of using internal grades is that they do not influence teachers' probability of obtaining the bonus. However, internal grades are not awarded using a uniform criterion as standardized tests are. Since each school may have its own grading standards, differences in internal grades may not necessarily reflect differences in learning across schools, and so we restrict our comparison to students from the same school to control for school-specific grading standards.[37] Identifying the causal effect of BE requires that internal grades capture changes in learning across different grades within the same school. There are two reasons why this might not hold. Firstly, teachers' grading biases may lead to a weak correspondence between changes in learning and changes in internal grades. This concern is underscored 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 (Burgess and Greaves, 2013; Botelho et al., 2015), and caste (Hanna and Linden, 2012), others find no such disparities (Newstead and Dennis, 1990; Baird, 1998; Van Ewijk, 2011). Secondly, if teachers grade on a relative basis (e.g., the worst 10% always fails, or the top 10% always gets the highest grade), internal grades may not be able to capture absolute changes in student learning. Below we present evidence that our identification is not invalidated by these two issues.

If there is no bias in internal grading, standardized test scores and internal grades should broadly follow the same patterns when comparing students from the same school. Unfortunately, data disclosure policies do not allow us to match a student's internal grades to his/her results in the ECE. However, the Peruvian Ministry of Education provides a database with individual ECE test

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

[37]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 the 8th and 9th grade, as shown in Table 7, unobserved teacher characteristics are unlikely to bias our estimates.

scores, gender, an index of socioeconomic status,[38] and anonymized school identifiers (which differ from the school identifiers in SIAGIE). Controlling for school fixed effects, 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 (Panel A of Table A.5). Analogous regressions with 8th grade students' internal grades as the dependent variable (Panel B of Table A.5) show that the within-school correlation between students' achievement and their gender and socioeconomic status is qualitatively similar. It thus appears as if internal grades measure student learning in a manner roughly consistent with standardized test scores, providing evidence against the concern that internal grades are biased in terms of observable student characteristics.

Having established that internal grades are consistent with standardized test scores, we now examine whether grading on a curve is common in Peruvian secondary schools. If teachers were assigning grades on a relative basis, we would expect two different classes taught by the same teacher to have a very similar grade distribution. Our database on teachers shows that on average, 8th grade teachers from schools with only two classes teach in 93% of them, meaning that the teachers handing out the grades are practically the same across classes. We restrict our sample to 8th graders in schools with just two 8th grade groups in 2014 (18% of our schools), and test whether math and language internal grades have a different mean and standard deviation across the two classes belonging to the same school. With a significance level of 10%, we reject the null hypothesis of equality in means across both groups in math in 23% of schools. In the case of language, we reject that the average grades across classes are the same in 32% of schools. The average difference in means across groups is 0.66 and 0.75 in math and language, approximately one third of a standard deviation. An F-test for the equality of variances shows that in 23% and 20% of schools, we can reject the null hypothesis that the distribution of math and language grades

[38]The socioeconomic status index is constructed using the years of schooling of the students' parents, the ratio of individuals to rooms in their household, indicators for whether the family owns durable goods and several types of books, the material of which the walls, floors and roofs are made, and whether the house has running water, electricity, internet, sewage, and a phone connection. These characteristics were obtained from a questionnaire completed by students during the evaluation. The socioeconomic status index, with mean 0 and standard deviation 1, was computed using principal-component analysis.

has the same standard deviation.[39] The difference in means and standard deviations and their corresponding p-values are depicted in Appendix Figure A.1. All in all, this evidence points to the fact that grading on a curve is not the norm in Peruvian high schools.

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 considers a series of potential explanations as to why the program had a null effect and provides suggestive evidence on how plausible each of these explanations are.

7.1 Teachers Did Not Know or Understand the Program

One explanation for the null effect we find is that teachers did not hear about BE, or did not understand how it worked. We argue that it is unlikely that our results are driven by these factors. Along with the launch of the 2015 edition of the BE program, the Ministry of Education headed a diffusion campaign in 2015, making it likely that secondary school teachers were informed about the program. The fact that the principal and every teacher are paid if their school wins generates strong incentives for people working in the same establishment to inform each other about the BE. In our teacher survey, 64% of those who taught math or language in 8th 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 (respondents could select more than one option).

The system by which schools were scored under the BE was not overly complex.[40] It should have been relatively clear from a teacher's perspective that the main component of the BE score is the school's average ECE score, because performance in the ECE test was also the main component

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

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

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.[41] This is broadly confirmed by our survey, in which 64% of math or language 8th grade teachers who knew about BE in 2015 responded that the ECE test scores were the most important or second most important component of the BE score.

We take advantage of the fact that almost half of the schools in our sample share their building with a primary school that had previously participated in the BE, and that 13% of the schools operate in the same building as a primary school BE winner, and test for heterogeneous effects across both dimensions. 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 grades in either of these groups of schools, as shown in columns (3), (4), (8) and (9) of Table 9. Thus, it is unlikely that the BE had no impact because of a lack of awareness of its existence or its rules.

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 detailed in Section 2, the average BE bonus represents approximately 24% of a monthly salary, and is thus 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.[42] Moreover, in the subsample of 8th grade math or language teachers who responded to our survey, 42% correctly identified the bonus amount or thought that it was larger.[43]

[41] 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

[42] The average bonus in Muralidharan and Sundararaman (2011) was around 35% of a monthly salary, wheres prizes were in-kind in Glewwe et al. (2010), and had an average value ranging between 12% and 21% of a teachers' monthly wage. The Chilean program studied by Contreras and Rau (2012) awarded an average bonus of 10% of a monthly salary. 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.

[43] However, only 30% of those teachers that knew about the program thought it was larger or adequate. This estimate may be downward biased, as many teachers took the survey as an opportunity to complain about their low salaries.

If the bonus was not large enough to incentivize the average teacher, we would perhaps find a positive effect in schools in which teachers' pay is relatively lower. As shown in columns (1) and (6) of Table 9, we do not find any heterogeneity by teachers' average salary in 2015.[44,45] 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 that the program had no distinguishable effect on students' math and language grades.

7.3 Group Incentives Do Not Work

Under a collective incentive scheme such as BE, the power of the incentive in promoting higher teacher effort is low when the group of incentivized teachers in a school is large. Teachers will have incentives to free-ride on their colleagues' effort, since the mapping between a teacher's actions and the probability of winning the bonus is weak when the number of incentivized teachers is high (Holmstrom, 1982). In addition, while collective incentives have the benefit of inducing higher cooperation and monitoring among teachers (Kandel and Lazear, 1992; Kandori, 1992), this might be hard to achieve when the group of teachers is very large. In 2013-2015, the average secondary school in our sample had two groups of 8th graders, meaning that the average school had four or less incentivized teachers (i.e., teaching math or language in 8th grade). This figure is comparable to other studies finding positive effects. In Muralidharan and Sundararaman (2011) and Glewwe et al. (2010), for example, the average schools has three and six incentivized teachers, respectively.

As shown in column (7) of Table 5, we do not find any differential effects by the number of 8th grade groups. If we break the results down even more, as shown in column (2) of Table 9, we do find that BE had a statistically significant but small effect on the math grades of students in schools with only one class per grade (covering 21% of students in 60% of schools). More specifically, the teacher incentive increases math grades by 0.022 SD.[46] However, since these effects are

[44]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.

[45]It should be noted, however, that teachers with a lower salary tend to be younger, have less experience, and are more likely to be contract teachers as compared to those with a higher salary.

[46]The sum of the 8^{th} *Grade* $\times$ *Post* and 8^{th} *Grade* $\times$ *Post* $\times$ *One Class* coefficients yields a total effect of 0.022

substantially smaller than those found in other studies in the literature, the fact that the incentive faced by teachers under the BE is collective does not seem to be one of the main reasons behind the program's null effect.

7.4 Teachers Only Focused on Improving Standardized Test Scores

As discussed in Section 3, teacher incentive programs may not result in higher learning if teachers focus their efforts on short-term strategies aimed solely at increasing standardized test scores. One explanation for why we do not observe any improvement in internal grades is that teachers may have reacted to the incentive by targeting topics likely to appear in the ECE, coaching students on test-taking strategies, or even cheating. However, there are reasons for believing that teachers could not engage in this type of behavior. Firstly, the implementation was carried out by independent officials who were trained and responded directly to the Peruvian National Statistics Institute. Teachers were not allowed to be in the room at any moment during the exam, and were not responsible for its correction, making cheating highly unlikely. Since students had no personal stake in this exam, there was no incentive to cheat on their part. Secondly, because the ECE exam is designed to capture a wide range of skills,[47] teachers could not benefit from 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 not 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 unlikely in this setting (Neal, 2011).

SD, with a p-value of 0.002.

[47] 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.

7.5 Teachers Did Not Know How to Improve Standardized Test Scores

Another explanation for the null effect we uncover is that teachers may not have known how to improve their students' performance in the standardized tests tied to the bonus. Unlike the intervention in Muralidharan and Sundararaman (2011) where teachers receive detailed feedback on their students' performance in baseline tests, Peruvian teachers receive no such information, as their students only take the ECE at the end of the school year. This is to be expected in a national program, where the cost of testing students twice a year is prohibitive. The lack of teacher knowledge on how to improve their students' performance in the ECE may have acted as a binding constraint, limiting the success of BE.[48] This may have been aggravated by the fact that 2015 was the first year in which the ECE test was implemented in secondary schools, providing even more uncertainty about how to raise ECE test scores.

Even if teachers knew how to equip their students with the skills needed to obtain high ECE scores, they may have encountered difficulties in passing on the incentive to their students. 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. This means that students have little to no incentive to exert effort into preparing for and taking these tests.[49] If teachers anticipate that their actions will only marginally impact their students' ECE scores, they may also be discouraged from exerting more effort.[50] The results from Behrman et al. (2015) provide suggestive evidence on the possibility that incentivizing teachers on their

[48] A recent study in Argentina shows that providing schools with diagnostic and feedback on their students' performance can improve classroom practices and student outcomes (de Hoyos et al., 2017). Although Muralidharan and Sundararaman (2010) find no impact of giving teachers in rural India information on their students' performance and information on how to use this feedback, this intervention increased student performance when it was paired with monetary incentives to teachers (the treatment from Muralidharan and Sundararaman, 2011). It is thus possible that teachers' knowledge and incentives are complementary inputs.

[49] A few recent studies show moderate levels of test-taking effort in PISA evaluations (also low-stakes for students). Borgonovi and Biecek (2016) and Zamarro et al. (2016) find that students are almost 14 percentage points less likely to answer a question correctly if it appears in the last ten items on the test, rather than the first ten items. Moreover, this decline in student effort is higher for lstudents in ow-performing countries.

[50] An additional explanation is that schools may not have had enough time to increase their students' learning in a meaningful way. Even though the program was first announced at the end of 2014, it was only confirmed in July 2015, four months before the November 2015 ECE. However, the study by Lavy (2002) finds a positive impact in a similar time span.

students' performance may not be effective unless students have a stake as well.[51]

8 Conclusion

We examine the short-run impact of a scaled-up teacher pay-for-performance program (*Bono Escuela*) implemented in 2015 in 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 with certain characteristics. We stipulate that the lack of increase in student learning may have been triggered by certain aspects of the evaluation linked to the bonus (students' low stakes and teachers' inexperience with it). These factors may have discouraged teachers from exerting greater effort.

Overall, the results from our study suggest that successfully scaling up teacher pay-for-performance requires a deeper understanding about the role played by the characteristics of these programs in their success. In particular, our findings raise the question of whether teacher incentives should be paired with teacher feedback and training in order for these programs to work. This paper also points to the fact that the type of exam being incentivized, and particularly the stakes that the students have in it, may be important for teacher incentives to be able to raise the levels of 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.

[51]The evidence on this point is only suggestive because, as compared to the treatment arms in which only teachers or students were incentivized, the potential reward was larger when both students and teachers were incentivized.

References

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.

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.

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 Alejandro J Ganimian, "The Barking Dog that Bites: Test Score Volatility and School Rankings in Punjab, Pakistan," *International Journal of Educational Development*, 2016, *49*, 31–54.

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

Behrman, Jere R, 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.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur, "Experimental Evidence on Scaling Up Education Reforms in Kenya," *Journal of Public Economics*, 2018, *168*, 1–20.

Borgonovi, Francesca and Przemyslaw Biecek, "An International Comparison of Students' Ability to Endure Fatigue and Maintain Motivation During a Low-Stakes Test," *Learning and Individual Differences*, 2016, *49*, 128–137.

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.

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.

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

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.

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.

de Hoyos, Rafael, Alejandro J Ganimian, and Peter A Holland, "Teaching with the Test: Experimental Evidence on Diagnostic Feedback and Capacity Building for Public Schools in Argentina," 2017. World Bank Working Paper 8261.

Deaton, Angus and Nancy Cartwright, "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science & Medicine*, 2018, *210*, 2–21.

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

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.

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

Gilligan, Daniel O, Naureen Karachiwalla, Ibrahim Kasirye, Adrienne Lucas, and Derek A Neal, "Educator Incentives and Educational Triage in Rural Primary Schools," 2018. NBER Working Paper 24911.

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

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.

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.

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.

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.

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.

Kerwin, Jason T and Rebecca Thornton, "Making the Grade: Understanding What Works for Teaching Literacy in Rural Uganda," 2015.

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.

Lazear, Edward P and Sherwin Rosen, “Rank-Order Tournaments as Optimum Labor Contracts,” *Journal of Political Economy*, 1981, *89* (5), 841–864.

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.

_ **, Pilar Romaguera, and Miguel Urquiola**, “Socioeconomic Status or Noise? Tradeoffs in the Generation of School Quality Information,” *Journal of Development Economics*, 2007, *84* (1), 61–75.

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.

Rivkin, Steven G, Eric A Hanushek, and John F Kain, “Teachers, Schools, and Academic Achievement,” *Econometrica*, 2005, *73* (2), 417–458.

Rockoff, Jonah E, "The Impact of Individual Teachers on Student Achievement: Evidence From Panel Data," *American Economic Review*, 2004, *94* (2), 247–252.

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.

Staiger, Douglas O and Jonah E Rockoff, "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives*, 2010, *24* (3), 97–118.

Zamarro, Gema, Collin Hitt, and Ildefonso Mendez, "When Students Don't Care: Reexamining International Differences in Achievement and Non-Cognitive Skills," 2016. EDRE Working Paper 2016-18.

Figure 1: Timing of BE in Secondary Schools

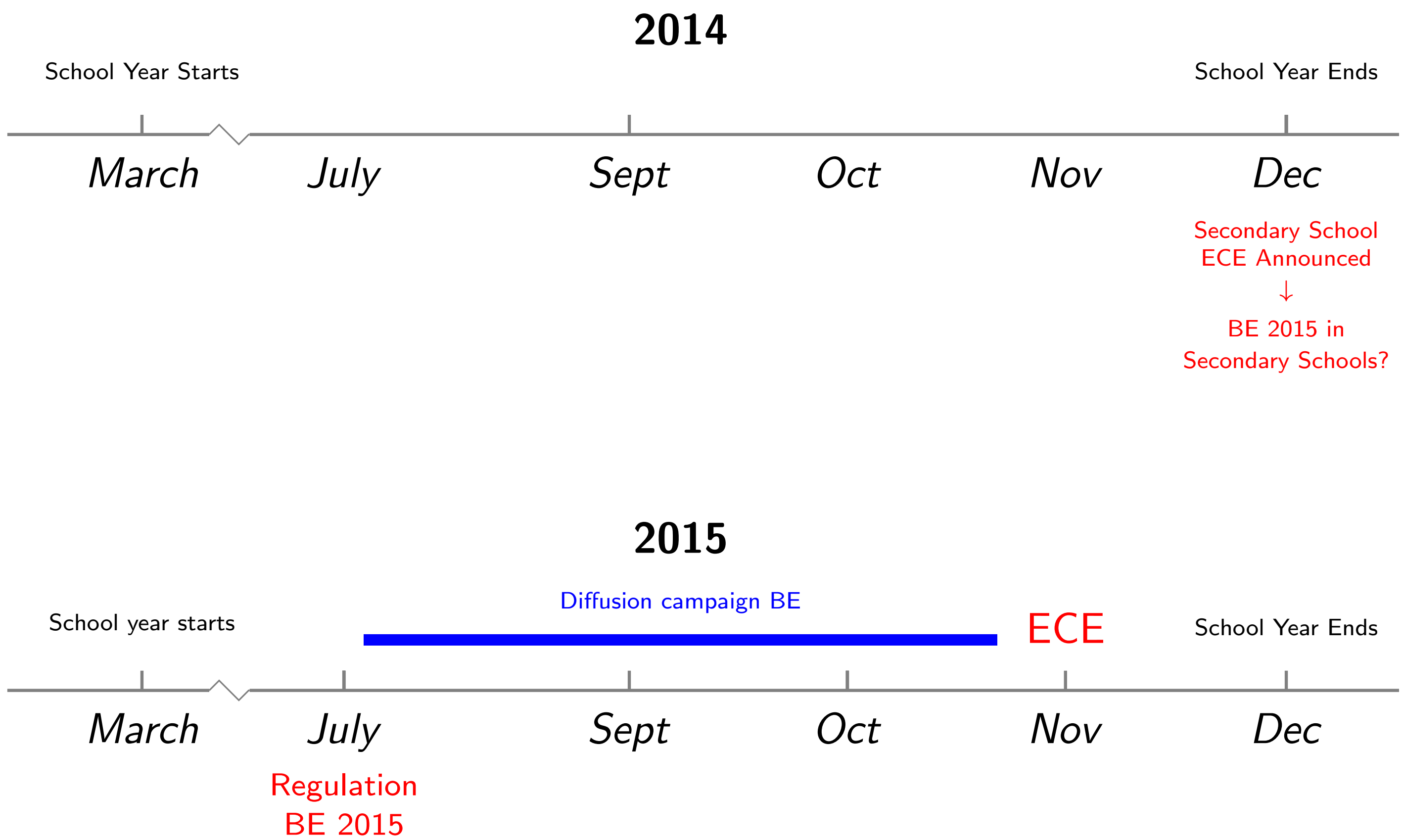
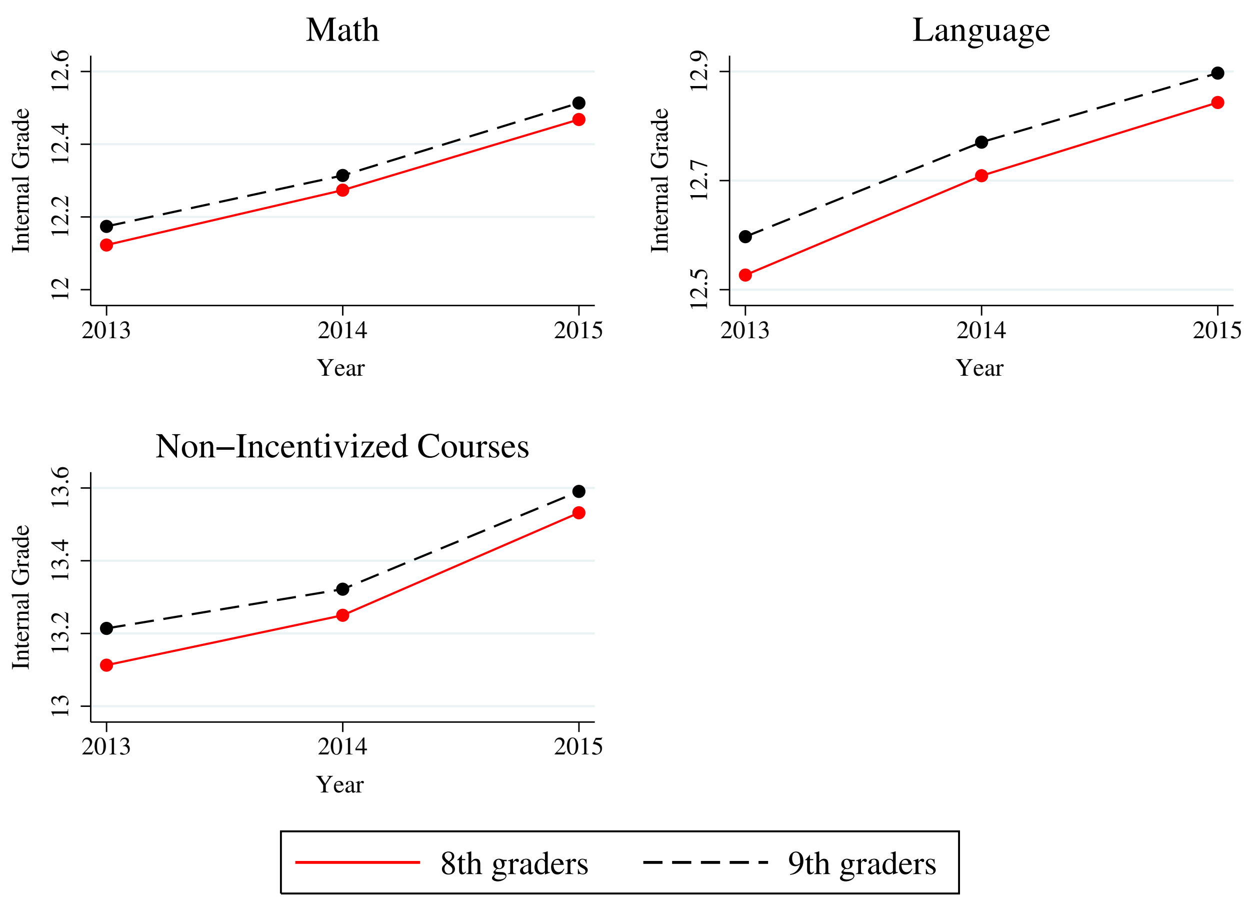
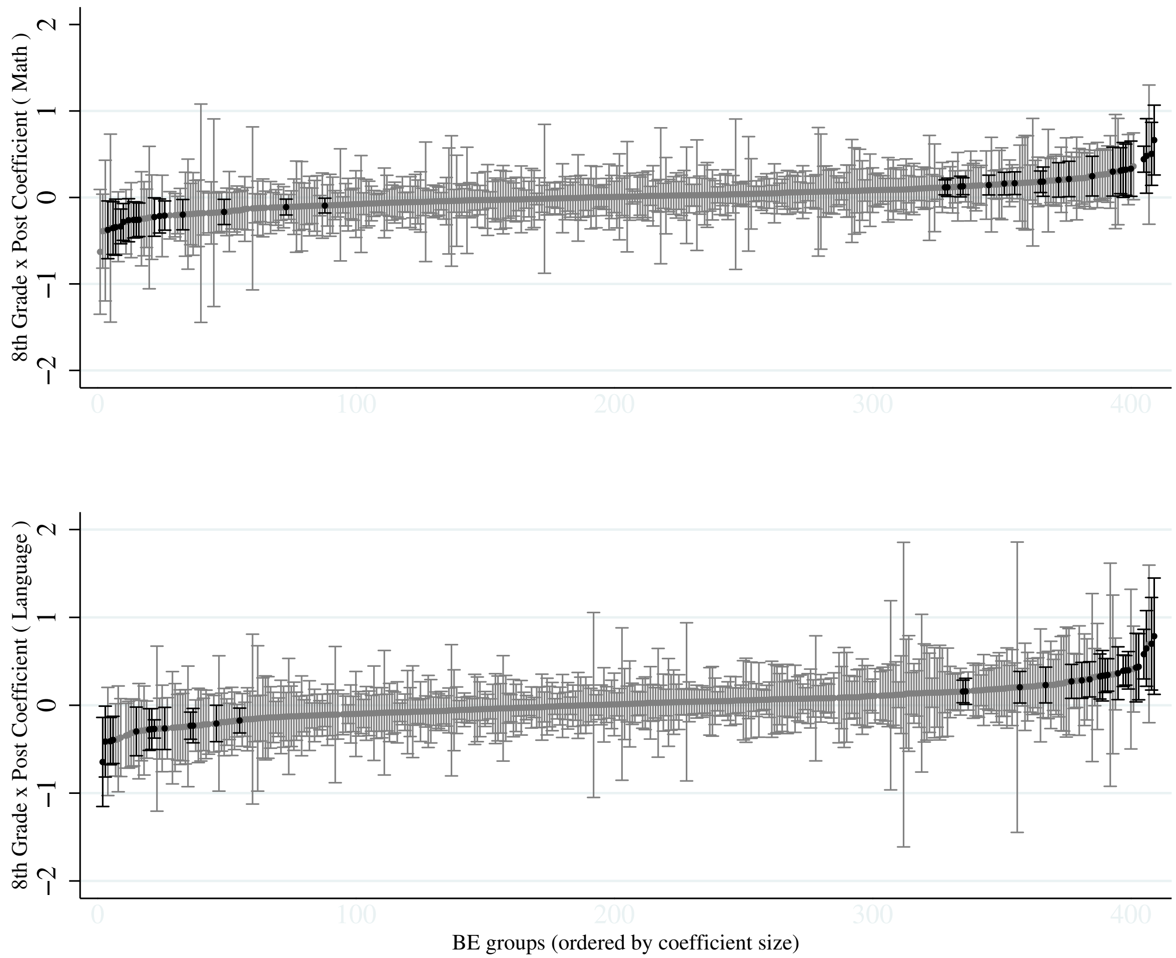


Figure 2: Trend in Average Internal Grades for 8th and 9th Graders



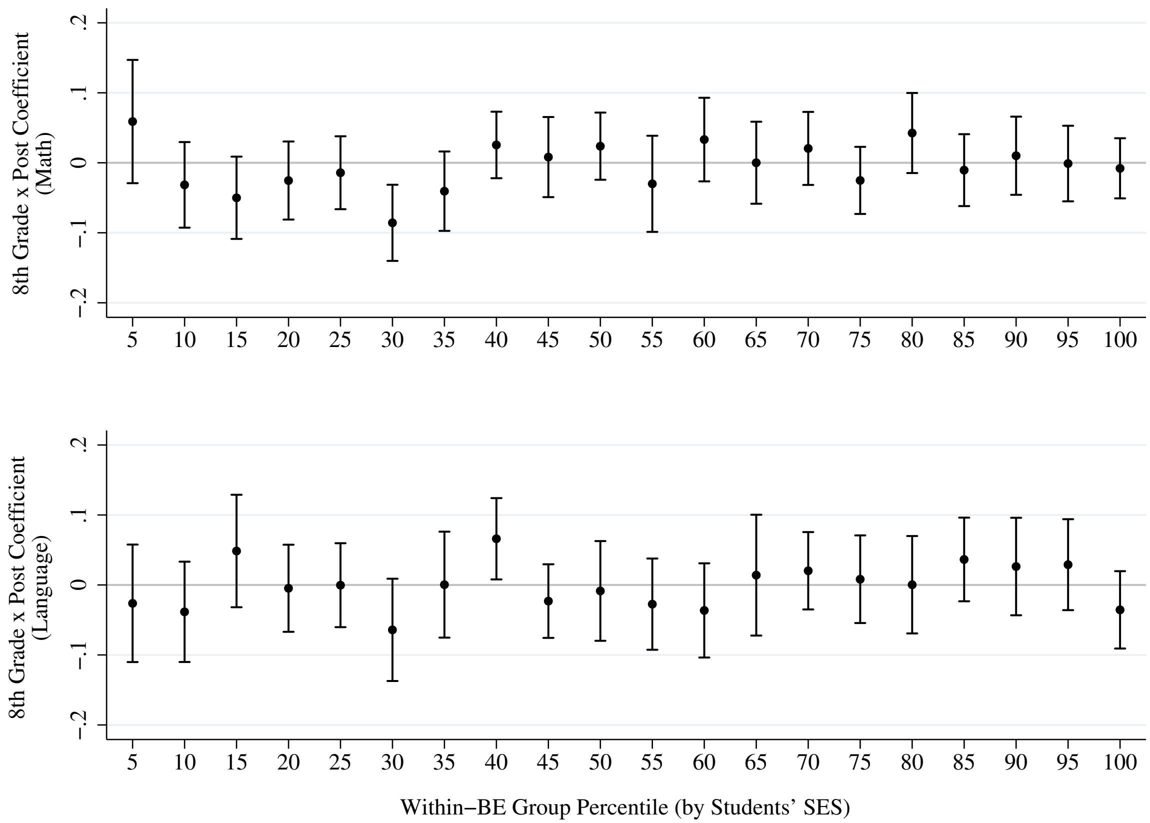
Notes: The sample includes all 8th and 9th grade students attending public secondary school in 2013-2015, in public schools which were eligible for taking the 2015 ECE standardized test and which are registered in the Ministry of Education's SIAGIE administrative system. The figures plot the average of all 8th and 9th 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.

Figure 3: Effect of Teacher Incentives on Students' Math and Language Internal Grades in each BE Group



Notes: The sample includes all 8th and 9th 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. Coefficients and confidence intervals in black are statistically significant at the 5% level, whereas those in grey are not. BE groups in both figures are ordered by coefficient size, and the ordering is separately done in each figure.

Figure 4: Heterogeneous Effect of Teacher Incentives on Students' Internal Grades by Schools' Socioeconomic Status (SES) Rank



Notes: The sample includes all 8th and 9th grade students attending public secondary school in 2013-2015, in public schools eligible to take 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 subject and year. We construct a SES index (taking values 0-3) for each 8th grade 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 groups according to their ranking in terms of this measure within their BE group. The figures plot the coefficient and 95% confidence interval for the interaction of each percentile dummy and the 8th x Post dummy from our baseline regressions.

Table 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.

Table 2: Summary Statistics for 8th and 9th Graders

	8th Grade		9th Grade	
	Mean	Std. Dev	Mean	Std. Dev
Final Grade (0-20)				
Math	12.29	2.18	12.33	2.18
Language	12.69	2.07	12.76	2.07
Other courses - average	13.30	1.60	13.38	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	0.05	0.23	0.04	0.20
Foreigner	0.00	0.05	0.00	0.05
Spanish is native tongue	0.85	0.36	0.84	0.37
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	0.98	0.15	0.97	0.16
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	1,122,064		1,052,043	
Grade/School Characteristics				
Rural	0.39	0.49	0.39	0.49
Number of classes	2.05	1.94	1.97	1.85
Teacher-pupil ratio	20.09	8.82	19.37	8.85
Number of teachers	11.00	6.52	11.11	6.68
Share of teachers instructing the other grade	0.84	0.20	0.83	0.20
Average age of teachers	41.52	17.08	41.46	17.15
Share of male teachers	0.59	0.20	0.60	0.20
Number of school-year observations	23,314		23,328	

Notes: The sample includes all 8th and 9th grade students attending public secondary school in 2013-2015, in public schools eligible to take the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. We exclude students for whom we have no grades and/or no individual controls (0.8%). Since teacher data is missing for a small subsample of schools, the number of grade-observations for teacher characteristics is 23,314 and 23,328 in 8th and 9th grades (instead of 23,634). *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 8th (9th) teachers also teaching 9th (8th) grade in the same school.

Table 3: Effect of Teacher Incentives on Math and Language Internal Grades

	Math			Language		
	(1)	(2)	(3)	(4)	(5)	(6)
8^{th} Grade x Post	0.001	-0.001	-0.003	0.006	0.004	0.001
	(0.007)	(0.006)	(0.006)	(0.008)	(0.008)	(0.008)
8^{th} Grade	-0.021***	-0.015***	-0.007*	-0.032***	-0.025***	-0.015***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Repeated last year			-0.570***			-0.623***
			(0.006)			(0.007)
Male			0.116***			0.340***
			(0.003)			(0.003)
Foreigner			0.044**			0.089***
			(0.018)			(0.018)
Spanish is native tongue			0.180***			0.183***
			(0.007)			(0.007)
Has a disability			-0.251***			-0.248***
			(0.013)			(0.014)
Father is alive			0.069***			0.069***
			(0.003)			(0.003)
Mother is alive			0.040***			0.048***
			(0.006)			(0.006)
Father lives in HH			0.029***			0.022***
			(0.002)			(0.002)
Mother lives in HH			0.013***			0.018***
			(0.003)			(0.002)
Observations	2,183,464	2,183,464	2,174,162	2,183,450	2,183,450	2,174,147
R^2	0.000	0.072	0.092	0.000	0.088	0.136
Year FE	✓	✓	✓	✓	✓	✓
School FE		✓	✓		✓	✓
Individual Controls			✓			✓

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible to take 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 subject and year. 8^{th} *Grade* is a dummy for whether the student is in 8^{th} grade, and *Post* is a dummy for the year 2015. Standard errors clustered by school are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 4: Effect of Teacher Incentives on Internal Grades in Non-Incentivized Courses

	Non-Incentivized Courses		
	(1)	(2)	(3)
8^{th} Grade x Post	0.015***	0.014***	0.011***
	(0.004)	(0.004)	(0.004)
8^{th} Grade	-0.042***	-0.036***	-0.027***
	(0.003)	(0.003)	(0.002)
Repeated last year			-0.599***
			(0.006)
Male			0.288***
			(0.002)
Foreigner			0.046***
			(0.014)
Spanish is native tongue			0.121***
			(0.005)
Has a disability			-0.225***
			(0.012)
Father is alive			0.064***
			(0.003)
Mother is alive			0.043***
			(0.005)
Father lives in HH			0.024***
			(0.002)
Mother lives in HH			0.012***
			(0.002)
Observations	2,183,628	2,183,628	2,174,321
R^2	0.001	0.124	0.189
Year FE	✓	✓	✓
School FE		✓	✓
Individual Controls			✓

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible to take 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 subject and 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 student is in 8^{th} grade, and *Post* is a dummy for the year 2015. Standard errors clustered by school are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Heterogeneous Effect of Teacher Incentives on Math and Language Internal Grades

	Male	Spanish Speaker	Parents' Education	Repeated	Lagged Grade	Ln (Enrollment)	Num. Classes	Rural
Panel A: Math Grades								
8^{th} Grade x Post	-0.007	-0.004	-0.005	-0.002	-0.002	0.030	-0.002	-0.005
	(0.008)	(0.012)	(0.007)	(0.007)	(0.008)	(0.023)	(0.009)	(0.007)
8^{th} Grade x Post x Covariate	0.009	0.002	-0.002	-0.015	0.007	-0.007	-0.000	0.017
	(0.009)	(0.013)	(0.005)	(0.016)	(0.005)	(0.006)	(0.002)	(0.012)
Observations	2,174,162	2,174,162	1,918,272	2,174,162	1,424,563	2,174,162	2,174,162	2,174,162
R^2	0.092	0.092	0.100	0.092	0.446	0.093	0.092	0.092
Panel B: Language Grades								
8^{th} Grade x Post	0.007	0.002	-0.002	0.002	0.004	-0.015	-0.007	0.002
	(0.009)	(0.014)	(0.009)	(0.008)	(0.009)	(0.028)	(0.010)	(0.009)
8^{th} Grade x Post x Covariate	-0.012	-0.001	0.000	-0.003	0.001	0.004	0.002	-0.010
	(0.010)	(0.015)	(0.006)	(0.016)	(0.005)	(0.007)	(0.002)	(0.014)
Observations	2,174,147	2,174,147	1,918,259	2,174,147	1,424,507	2,174,147	2,174,147	2,174,147
R^2	0.136	0.136	0.143	0.136	0.421	0.137	0.136	0.136

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible to take the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. Heterogeneities by *Lagged Grade* exclude students from 2013, for which the 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 subject and year. 8^{th} *Grade* is a dummy for whether the student 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 report only two coefficients for exposition purposes. *Spanish Speaker* is a dummy for whether the student's first language is Spanish, and *Parents' Education* 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 achieved that level of education. *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 number of students enrolled in that year and grade (in ln). *Num. Classes* is the number of classes in the student's grade and year, and *Rural* is a dummy for whether the school is in a rural area. Standard errors clustered by school are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 6: Test for Parallel Trends in Students' Internal Grades

	Math	Language	Non-Incentivized Courses
8^{th} Grade x Post	-0.001	0.002	0.017***
	(0.007)	(0.009)	(0.004)
8^{th} Grade x 2014	0.004	0.003	0.011***
	(0.007)	(0.009)	(0.004)
8^{th} Grade	-0.009*	-0.016**	-0.033***
	(0.005)	(0.006)	(0.003)
Observations	2,174,162	2,174,147	2,174,321
R^2	0.092	0.136	0.189

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible to take 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 subject and 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 student 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 are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 7: Effect of Teacher Incentives on Teacher Characteristics

	Average Age	Share Male	Average Number of Classes	Average Number of Schools	Share New to School	Share New to Grade
8^{th} Grade x Post	-0.077	0.000	-0.004	-0.001	-0.000	0.000
	(0.105)	(0.001)	(0.005)	(0.001)	(0.001)	(0.002)
8^{th} Grade	0.082	-0.006***	0.053***	-0.002***	0.001	0.002
	(0.082)	(0.001)	(0.003)	(0.001)	(0.001)	(0.002)
Observations	46,642	46,642	46,642	46,642	31,079	31,079
R^2	0.476	0.748	0.917	0.615	0.796	0.640
Dependent Variable Mean	41.490	0.594	1.549	1.134	0.458	0.310

Notes: The sample includes all public secondary schools in 2013-2015, in public schools eligible to take 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 *Share Male* is the share of teachers in that grade that are male. *Average Number of Classes* is the average number of classes taught by teachers in that grade, and *Average Number of Schools* is the average number of schools in which the teacher works in that year. *Share New to School* and *Share New to Grade* are the proportion of teachers who are new to the school and grade, respectively. All regressions include school fixed effects and year fixed effects. 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 are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 8: Heterogeneity by Overlap of 8th and 9th Grade Teachers

	Math			Language		
	Low	Med	High	Low	Med	High
8th Grade x Post	-0.023	0.003	0.009	0.018	-0.008	-0.006
	(0.014)	(0.012)	(0.007)	(0.018)	(0.014)	(0.009)
8th Grade	0.001	-0.005	-0.016***	-0.015	-0.014*	-0.016***
	(0.008)	(0.007)	(0.004)	(0.011)	(0.008)	(0.006)
Repeated last year	-0.593***	-0.546***	-0.572***	-0.635***	-0.597***	-0.639***
	(0.014)	(0.008)	(0.008)	(0.015)	(0.009)	(0.009)
Male	0.113***	0.135***	0.103***	0.368***	0.369***	0.293***
	(0.006)	(0.005)	(0.004)	(0.007)	(0.006)	(0.004)
Foreigner	0.043	0.025	0.083**	0.076***	0.080***	0.145***
	(0.027)	(0.029)	(0.042)	(0.027)	(0.029)	(0.041)
Spanish is native tongue	0.164***	0.181***	0.191***	0.163***	0.179***	0.202***
	(0.015)	(0.011)	(0.009)	(0.014)	(0.013)	(0.009)
Has a disability	-0.243***	-0.207***	-0.297***	-0.222***	-0.206***	-0.305***
	(0.030)	(0.023)	(0.019)	(0.026)	(0.026)	(0.022)
Father is alive	0.069***	0.075***	0.064***	0.074***	0.067***	0.065***
	(0.007)	(0.005)	(0.005)	(0.007)	(0.006)	(0.005)
Mother is alive	0.027*	0.038***	0.052***	0.032***	0.044***	0.063***
	(0.014)	(0.010)	(0.008)	(0.012)	(0.010)	(0.009)
Father lives in HH	0.036***	0.034***	0.016***	0.028***	0.025***	0.012***
	(0.005)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)
Mother lives in HH	0.010**	0.012***	0.015***	0.014***	0.019***	0.020***
	(0.005)	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)
Observations	671,155	743,433	758,072	671,146	743,435	758,065
R^2	0.074	0.083	0.122	0.119	0.126	0.165
Avg. share of teachers in both grades	0.347	0.608	0.907	0.347	0.608	0.907

Notes: The sample includes all 8th and 9th grade students attending public secondary school in 2013-2015, in public schools eligible to take 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 8th and 9th grade teachers in 2015. Overlap between 8th and 9th grade teachers is the % of teachers in 8th grade also instructing 9th grade. The dependent variables are students' internal grades in math and language, standardized by subject and year. *8th Grade* is a dummy for whether the student is in 8th grade, and *Post* is a dummy for the year 2015. All regressions include school fixed effects and year fixed effects. Standard errors clustered by school are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 9: Heterogeneous Effect of Teacher Incentives by Average Salary, Number of Classes, and School's Experience with Primary School BE

	Math				Language			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
8^{th} Grade x Post	0.233	-0.010	-0.005	-0.005	0.361	-0.001	0.006	0.000
	(0.438)	(0.008)	(0.009)	(0.007)	(0.517)	(0.010)	(0.011)	(0.008)
8^{th} Grade x Post x Ln (Average Salary)	-0.032				-0.049			
	(0.060)				(0.071)			
8^{th} Grade x Post x One Group		0.032***				0.010		
		(0.011)				(0.013)		
8^{th} Grade x Post x BE Primary			0.004				-0.009	
			(0.013)				(0.016)	
8^{th} Grade x Post x BE Primary Winner				0.015				0.005
				(0.017)				(0.024)
Observations	2,143,420	2,174,162	2,174,162	2,174,162	2,143,405	2,174,147	2,174,147	2,1741,47
R^2	0.092	0.092	0.092	0.092	0.136	0.136	0.136	0.136
P-Value (sum of both coefficients = 0)	0.596	0.002	0.914	0.532	0.485	0.280	0.802	0.824

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible to take 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 subject and year. 8^{th} *Grade* is a dummy for whether the student 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. 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 are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

ONLINE APPENDIX

Appendix A Appendix Figures and Tables

Table A.1: Test for Parallel Trends Comparing Public and Private Schools

	Math	Language	Non-Incentivized Courses
Public x Post	0.124***	0.113***	0.156***
	(0.010)	(0.011)	(0.007)
Public x 2014	0.048***	0.069***	0.046***
	(0.009)	(0.010)	(0.006)
Repeated last year	-0.528***	-0.588***	-0.551***
	(0.006)	(0.006)	(0.005)
Male	0.108***	0.323***	0.272***
	(0.003)	(0.003)	(0.002)
Foreigner	0.042***	0.061***	0.064***
	(0.012)	(0.011)	(0.009)
Spanish is native tongue	0.197***	0.205***	0.132***
	(0.007)	(0.007)	(0.005)
Has a disability	-0.245***	-0.259***	-0.229***
	(0.015)	(0.015)	(0.013)
Father is alive	0.075***	0.077***	0.070***
	(0.004)	(0.004)	(0.003)
Mother is alive	0.031***	0.039***	0.038***
	(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	0.010***	0.015***	0.010***
	(0.002)	(0.002)	(0.002)
Observations	1,476,616	1,476,600	1,476,709
R^2	0.152	0.194	0.291

Notes: The sample includes all 8th grade students in 2013-2015, in public and private schools eligible to take 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 subject and 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 student 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 displayed since this variable is perfectly collinear with the corresponding school fixed effect. Standard errors clustered by school are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A.2: Online Survey Responses

	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
More attention paid 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
More attention paid 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%

Table A.3: Effect of Teacher Incentives 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 x Post	0.001	0.002	0.017*	0.012	0.011	0.019**	0.013	0.013	0.011
	(0.009)	(0.008)	(0.009)	(0.008)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)
8^{th} Grade	-0.035***	0.052***	-0.039***	-0.046***	-0.054***	-0.043***	-0.015**	-0.039***	-0.024***
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
Repeated last year	-0.618***	-0.569***	-0.601***	-0.572***	-0.595***	-0.587***	-0.615***	-0.611***	-0.621***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.007)	(0.007)
Male	0.341***	0.239***	0.262***	0.283***	0.343***	0.390***	0.076***	0.397***	0.260***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)
Foreigner	0.028*	0.055***	0.040**	0.164***	0.041**	0.028	0.059***	-0.012	0.004
	(0.016)	(0.017)	(0.019)	(0.018)	(0.019)	(0.017)	(0.015)	(0.017)	(0.017)
Spanish is native tongue	0.084***	0.155***	0.131***	0.152***	0.131***	0.144***	0.088***	0.101***	0.106***
	(0.006)	(0.007)	(0.008)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)
Has a disability	-0.179***	-0.253***	-0.221***	-0.272***	-0.230***	-0.232***	-0.254***	-0.179***	-0.210***
	(0.015)	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)	(0.014)	(0.016)	(0.016)
Father is alive	0.064***	0.066***	0.068***	0.070***	0.065***	0.068***	0.051***	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***	0.040***	0.039***	0.049***	0.036***	0.043***	0.046***	0.036***	0.041***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)
Father lives in HH	0.025***	0.028***	0.026***	0.022***	0.025***	0.025***	0.013***	0.024***	0.024***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Mother lives in HH	0.012***	0.012***	0.013***	0.013***	0.013***	0.014***	0.010***	0.015***	0.009***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	2,174,156	2,174,152	2,174,141	2,174,149	2,174,150	2,174,137	2,174,154	2,137,066	2,174,136
R^2	0.200	0.129	0.141	0.146	0.156	0.171	0.239	0.184	0.189

Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2013-2015, in public schools eligible to take 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 subject and year. *8^{th} Grade* is a dummy for whether the student is in 8^{th} grade, and *Post* is a dummy for the year 2015. All regressions include school fixed effects and year fixed effects. Standard errors clustered by school are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A.4: Non-Linear Heterogeneous Effects by Students' Lagged Grade

	Math	Language
Panel A: Lagged Grade Quartiles		
8^{th} Grade x Post	-0.002	0.001
	(0.009)	(0.011)
8^{th} Grade x Post x Q2	-0.007	0.010
	(0.009)	(0.010)
8^{th} Grade x Post x Q3	-0.011	0.007
	(0.010)	(0.011)
8^{th} Grade x Post x Q4	0.015	0.002
	(0.014)	(0.014)
Observations	1,424,563	1,424,507
R^2	0.399	0.391
P-value (sum of coefficients Q2)	0.345	0.303
P-value (sum of coefficients Q3)	0.195	0.464
P-value (sum of coefficients Q4)	0.298	0.790
Panel B: Lagged Grade Terciles		
8^{th} Grade x Post	-0.004	-0.005
	(0.009)	(0.010)
8^{th} Grade x Post x T2	-0.012	0.008
	(0.009)	(0.009)
8^{th} Grade x Post x T3	0.016	0.013
	(0.012)	(0.012)
Observations	1,424,563	1,424,507
R^2	0.364	0.368
P-value (sum of coefficients T2)	0.071	0.739
P-value (sum of coefficients T3)	0.327	0.521

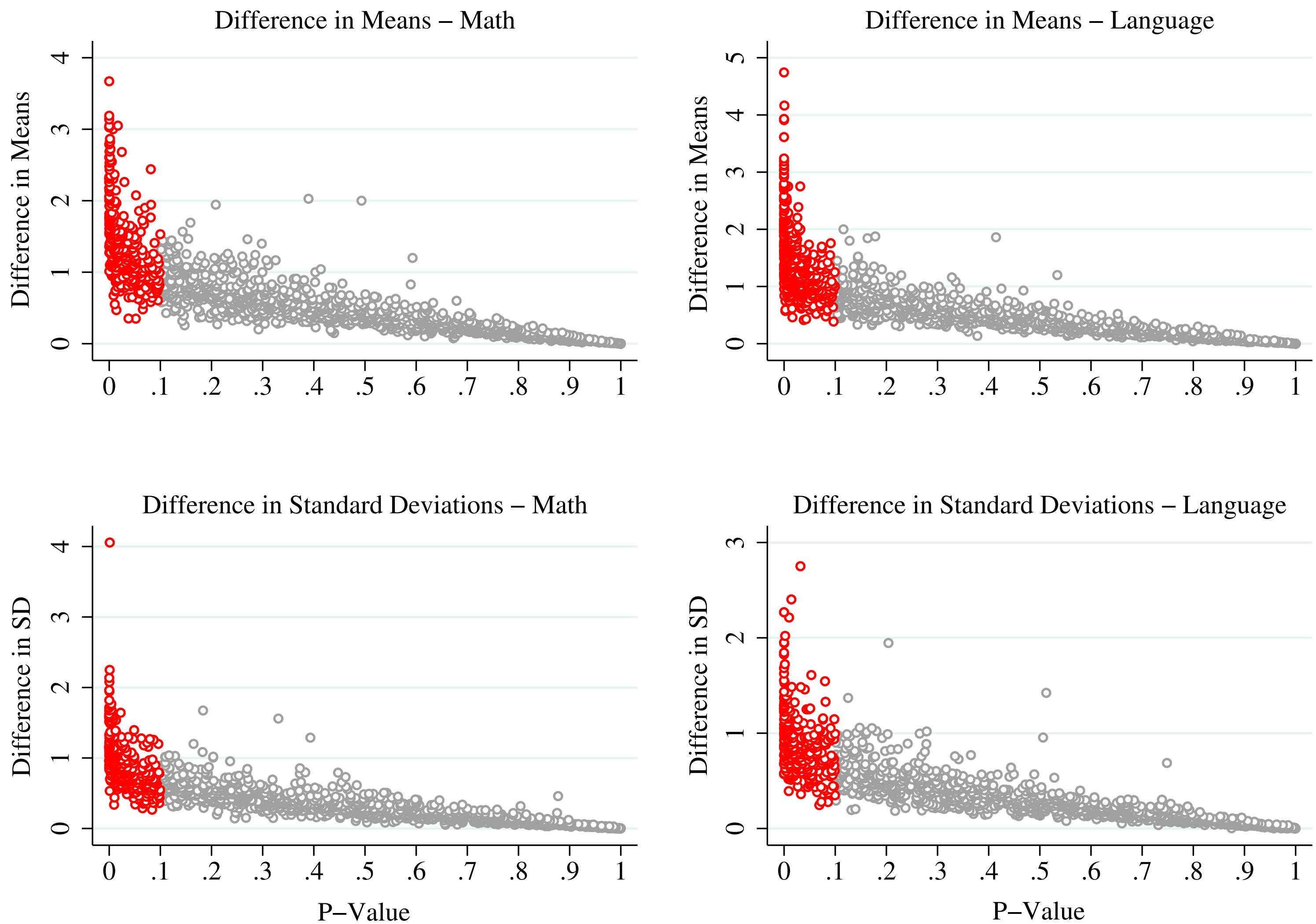
Notes: The sample includes all 8^{th} and 9^{th} grade students attending public secondary school in 2014-2015, in public schools eligible to take the 2015 ECE and registered in the Ministry of Education's SIAGIE administrative system. We exclude the year 2013 for which students' previous grade is unavailable. The dependent variables are students' internal grades in math and language, standardized by subject and 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 student 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 report only the triple interactions for exposition purposes. Standard errors clustered by school are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A.5: 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	354,429	354,547	354,529	354,547
R^2	0.020	0.216	0.015	0.254
Panel B: Internal Grades				
Spanish is native tongue	0.210***	-0.016***	0.224***	-0.010***
	(0.012)	(0.004)	(0.013)	(0.003)
Mother has high education	0.159***	-0.021***	0.166***	-0.014***
	(0.005)	(0.002)	(0.005)	(0.001)
Father has high education	0.129***	-0.016***	0.136***	-0.014***
	(0.004)	(0.002)	(0.005)	(0.001)
Male	0.139***	-0.035***	0.359***	-0.052***
	(0.005)	(0.002)	(0.005)	(0.001)
Observations	333,904	333,938	333,898	333,937
R^2	0.019	0.102	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, household assets and characteristics. All regressions include school fixed effects. Standard errors clustered by school are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Figure A.1: Variation in Internal Grades Across 8th Grade Classes in 2014



Notes: The top figures depict the difference in means and the corresponding p-value when testing whether the mean math and language internal grades differ across 8th grade classes from the same school in 2014, in every school with only two 8th 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.

Appendix B Teacher Survey

In conjunction with the Ministry of Education, we conducted an online survey to a subset of public secondary school teachers. According to SIAGIE, there were 123,669 public secondary school teachers in 2015. The Ministry of Education has the email address of 36,283 teachers (30%), all of whom received an email with the survey from the Ministry in October 2016, a few weeks before the winners of BE were announced. As in past editions of BE in primary schools, the bonus winners were announced at the end of the following school year (in November 2016).

The survey was anonymous, and teachers were told that its purpose was to collect information about teachers' perceptions and opinions about the BE program. 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 with non-respondents. We received a response from 3,406 teachers (9.4% response rate), approximately 2.8% of all public secondary school teachers. Given that a small group of teachers responded to the survey, our estimates may be biased due to self-selection. Yet the sign of this bias is not clear. Teachers that answered the questionnaire may be more motivated that the average public secondary school teacher. If this is the case, our estimates on the knowledge of teachers about the program will be upward biased. On the other hand, teachers that are dissatisfied with the educational system may take this survey as an opportunity to complain to the Ministry of Education[1] . This could result in a downward bias of our estimates. Moreover, as the survey was framed in the context of BE and sent by the Ministry of Education, the respondents may be subject to social desirability bias. This will result in an over-reporting of good behavior associated with the objectives of BE. Thus, given our selected sample and the potential bias in teachers' responses, we take the results from this survey only as suggestive evidence.

Teachers were asked what grades and subjects they taught in 2015, and their knowledge about the BE and its rules at that time. 64% of the surveyed teachers that taught math or language in 8^{th} grade in 2015 reported that they knew about the program's existence during the 2015 academic year. 57% of those teachers found out about the BE through the Ministry of Education or the school district authorities, 30% got the information from the news, and 35% from the school principal or other coworkers (respondents could select more than one option). The survey also confirmed that teachers understood the rules of the BE. 64% of math or language 8^{th} grade teachers who knew about BE in 2015 responded that the ECE test scores were the most important or second most important component of the BE score. We also asked teachers for their opinion on the size of the bonus. In the subsample of 8^{th} grade math or language teachers, 42% correctly identified the bonus

[1] After the survey email was sent, the Ministry of Education received several emails from secondary school teachers with complaints that were completely unrelated with the program, such as the poor infrastructure of their school, their salaries, or working conditions.

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 asked for 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 been due to the fact that the survey came from the Ministry of Education, and many teachers may have taken this as an opportunity to complain about their low salaries.[2]

In our online survey, we asked teachers whether they changed their pedagogical practices in 2015 as a result of BE, and separately asked about changes in pedagogy while teaching 8th grade as opposed to all other grades. As can be seen in Panel A of Table A.2, 8th grade teachers are five 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 (five percentage points), increased the difficulty of their classes (six percentage points), and increased the frequency of multiple choice examinations (nine percentage points). The same patterns hold when we restrict the analysis to teachers who taught math or language in the 8th grade and other grades, as seen in Panel B. However, 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.[3]

We tried to elicit teachers' perceptions of their school's ranking and probability of winning. One of the questions in our survey shows a hypothetical ranking of 20 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 improving the performance of their students (extra tutoring sessions, training sessions, etc.). In 47% of cases, math and language teachers in the 8th grade answered that their school would still be below the 80th percentile (i.e., would not win the bonus) after everyone changed their pedagogical practices. Finally we asked teachers for their opinions of students' motivation in the standardized test tied to the BE. When asked whether they thought students exerted any effort when taking the ECE test, 37% of survey respondents who taught math or language in 8th grade answered that they did not. We inquired about the reasons for this, teachers replied that ECE test scores do not affect final grades (51%), students are unmotivated

[2]In the open-ended part of this question, many teachers answered that their salaries are insufficient.

[3]In the study of Glewwe et al. (2010), for example, the survey of 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.

(47%), the test is too long (10%) or too difficult (8%), and students are not familiar with these types of evaluations (11%).