Performance Based Incentives for Learning in the Mexican Classroom

Andrew Christensen, MPA, Foundation Escalera

Brian Fuller, MPA, Foundation Escalera

Victor Steenbergen, MPA Candidate, London School of Economics

Alison Hamburg, MPA/MPH Candidate, Columbia University

Abstract

As the world continues to clamor for more coverage in education, the conversation on quality of learning remains comparatively quiet. This paper presents evidence of a policy solution that helps address the quality deficit in Mexico via the country's first randomized controlled trial of its kind. In 2010 the State Government of Chiapas began a two-year performance-based incentive program aimed at increasing learning outcomes. The program offered regular monetary and material rewards for high or improved achievement in the classroom. The program surveyed and tested nearly 8,000 students in 147 middle schools in predominantly indigenous and rural areas of Chiapas. Econometric analysis demonstrates that the program produced statistically significant gains for students in reading and mathematics, particularly for final year students. The study demonstrates benefits for low as well as high-performing students, as intended by design. The study also found significant increases in teacher attendance and parental support, which suggests that incentive programs can indirectly improve student learning by motivating teachers and parents. These results suggest that student incentives can be a viable policy tool in addressing the quality of learning.[1]

Keywords: Incentives, Randomized Controlled Trial (RCT), Mexico

[1] The authors wish to acknowledge our many project partners, especially the Benjamin Foundation, the W.K. Kellogg Foundation, Sra. Isabel Aguilera de Gutierrez of the Chiapas Ministry of Family Development (*DIF*), and Jorge Noriega Rocha of the Secretary of Education (*SEP*) for their kind support of this study. Many thanks go to the two capstone research teams from Columbia University's School of International and Public Affairs, and their advisor, Dr. Marta Vicarelli, for their hard work in making the program possible. Thank you also to Joseph Wales, Tobias Baedeker, Julia Brown, James McGibney and Fernando Morthera for their countless hours of work to make this project a reality.

1. Introduction

Over the past decade, interventions to improve schooling outcomes have grown to include not only supply-side reforms such as improvements in infrastructure, supplies or teacher quality, but also demand-side programs that provide monetary or non-monetary rewards to households conditioned on children's school enrollment, attendance or achievement (Berry 2011). The majority of these programs have focused on incentivizing student enrollment and attendance. Yet, research shows that interventions focused on increasing education coverage, while often successful in their primary goal, have not led to the expected improvements in student learning (Behrman, Parker, and Todd 2005; Schultz 2004; Ponce and Bedi 2010).

In recent years several studies have examined the effectiveness of providing incentives based on student achievement in the classroom (Kremer, Miguel, and Thornton 2004; Sharma 2010; Fryer 2010; Angrist and Lavy 2009; Berry 2011; Bishop 2004), working from the premise that providing direct, short-term rewards for student achievement may impact student learning by increasing students' extrinsic motivation to exert effort in school (Bishop 2004; Fryer 2010).

The Incentives to Excel Program (hereafter, the Star Program) is a performance-based incentives program aimed at increasing learning outcomes in marginalized secondary schools in the state of Chiapas, Mexico. In particular, the Star Program focuses on students in *telesecundarias*, middle schools that combine formal and distance education modalities and have characteristically lower learning outcomes than traditional *secundarias*. The Star Program began as a joint venture between the State Government of Chiapas and Foundation Escalera (hereafter, Escalera). The pilot year of the program took place during the 2010-11 academic year in highly marginalized rural zones. In the 2011-12 academic year the Star Program was reinstated and expanded to 147 schools (76 treatment and 71 control), reaching nearly 8,000 students. The

program offered monetary and non-monetary incentives to students on the basis of both overall grades and grade improvement. Learning outcomes in mathematics and reading were measured by a year-end exam administered by Escalera and the Chiapas Ministry of Education.

In our analysis of the 2011-12 program year, we find that treatment is associated with a rise in overall test scores of 0.237 standard deviations. We also find that the program had a larger impact on raising math scores than reading scores, and generally speaking the treatment effect is larger for male students than female students, with highest increases found in Grade 3 (U.S. Grade 9). In addition, we find an average treatment effect for the whole classroom, which suggests that the mix of incentives offered by the program is successful in motivating both low- and high-performing students. Interestingly, we also find substantial evidence that the program improves teacher attendance, teacher effort, and parental involvement, pointing to indirect channels through which the program raises test scores.

In the following sections we first review the literature on student incentives and provide a theoretical framework for our study and analysis. Next we provide an overview of the local context in which the program was administered, along with a description of our experimental design, data collection process, and econometric approach. Finally, we describe and interpret the study results, including the general treatment effect, outcomes broken down by gender and student grade, and the channels of influence for which we find significant evidence. We conclude by providing several policy recommendations and areas for future research.

2. Literature Review and Conceptual Framework

Interventions that reward student achievement are rooted in the theory that students possess both intrinsic motivation to study (the joy of learning), as well as extrinsic motivation to exert effort in school (based on perceived returns to education) (Fryer 2010; Bishop 2004).

According to Bishop (2004), students are rational actors who implicitly seek to maximize the benefits and minimize the costs of their efforts, and thus will choose the study effort level that maximizes the net benefit of their effort. If students lack sufficient intrinsic motivation to learn, dramatically discount the future or have pessimistic perceptions of future opportunities, or do not have accurate information on the returns to schooling, their internal cost-benefit calculation may lead them to not exert much effort in school (Bishop 2004; Fryer 2010). Incentive programs that provide direct, short-term rewards for student achievement thus seek to impact student effort by increasing students' extrinsic motivation to learn in such a way that is not contingent upon perceived future benefits of education (Fryer 2010; Bishop 2004).

Incentive types and methods

Studies of programs that reward student achievement have explored a number of factors that may impact the effectiveness of the intervention, including incentive delivery methods, types of incentives provided, eligibility for awards, and mechanisms for achieving rewards.

Incentive type and delivery. Most incentive programs have used cash rewards, while few studies have explored the effectiveness of non-monetary incentives. There has been some discussion as to whether cash transfers or non-monetary prizes are more effective in motivating students to learn. Berry (2011) argues that while monetary rewards are more tangible and thus more effective, children's parents may take ownership of cash prizes, which may lessen their motivation to achieve the award. In a study of academic incentives in India, for example, Berry found that a majority of students indicated that their parents had discretion over cash awards. In contrast, cash incentives may spark parents' desire for their child to win, resulting in increased parental involvement and, in turn, improved achievement (Kremer, Miguel, and Thornton 2004).

Non-monetary prizes allow students to maintain ownership of the award, and may hold

the additional value of exclusivity, especially in communities where such commodities are not commonplace (Berry 2011). Berry found that the effectiveness of different types of incentives depended on family background and initial academic performance: when initial academic performance was low and parents were less able to assist in their child's schooling behavior (generally poorer households with lower parental education), material incentives given directly to the child were more effective than cash incentives (Berry 2011). On the other hand, children with higher initial test scores performed better when offered a monetary prize (Berry 2011).

Award eligibility. Some incentive programs reward high achievement (Kremer, Miguel, and Thornton 2004; Leuven, Oosterbeek, and Klaauw 2003; Fryer 2010), while others reward academic improvement (Sharma 2010). Programs that reward high achievement run the risk of disincentivizing lower performers, who may cease to exert effort if they believe that they do not have a chance of winning (Leuven, Oosterbeek, and Klaauw 2003). In contrast, if award criteria are perceived as "too easy" by high performing students, the incentive may not be effective (Sharma 2010). Programs that reward high performing students may also have spillover effects for students with a lower chance of winning (Kremer, Miguel, and Thornton 2004).

Mechanisms for winning. Incentive programs employ different mechanisms for winning awards. In a tournament model (e.g., Kremer et al. 2004), only a certain percentage of top students are awarded. Yet, this model may disincentivize lower performing students who are not likely to win (Sharma, 2010). Blanket incentives, used in studies such as Berry (2009), Sharma (2010), Fryer (2010), and Angrist and Lavy (2009), reward all students who achieve a certain goal. Sharma (2010) found that for blanket incentives, lower performing students gain more than higher performing students. A lottery model, wherein students who meet a certain threshold are entered into a raffle, may be effective as they draw from the theory that individuals commonly

overestimate small probabilities (Haisley, Mostafa, and Loewenstein 2008).

Impact of incentive programs

While the results of incentive programs that reward student achievement have been mixed overall, almost all studies in both developed and developing countries find significant differences of treatment for at least some subsets of gender, subject, grade, native language, and initial academic performance. These findings bring to light several patterns in outcomes, including the salience of gender-based support, strategic effort allocation, and peer effects.

Several studies have found differences in results for male and female students. For example, Angrist and Lavy (2009) found that cash awards raised test performance among high school students in Israel, but only for female students. Higher performance among female students may reflect the theory that norms and behavior of young women are more supportive of learning than behavior of young men (Bishop 2004). Alternatively, women may have lower enrollment rates compared to men, and female students who remain in education may be systematically different than their peers (e.g., receive more parental support or have stronger motivation to learn) (Bishop 2004). In contrast, results from the Star Program pilot study found higher outcomes for male students. This may reflect Fryer's theory that to improve performance, student effort must be accompanied by complementary inputs (Fryer 2010); because parents may perceive boys' education to be more valuable, and teachers may believe girls are less able to perform, it may be harder for girls to increase their achievement. Moreover, Berry (2011) shows that students are incentivized to the extent that they are able to keep their prizes. If female students exercise less power over allocation of the household budget than their male counterparts, they might have a lower expected value for the same monetary incentive.

Several studies suggest that when faced with an achievement-based incentive, students

may strategically allocate their efforts toward those subjects in which they most easily improve (Sharma 2010; Bettinger 2010). For example in Nepal, Sharma found increases in academic performance, but only for "soft" subjects such as physical education, and the increases were larger for lower performing students. In a study of primary schools in the United States, Bettinger found significant results but only in mathematics, suggesting a similar pattern.

Finally, several studies of student incentive programs point to positive outcomes for peers and teachers, as predicted by Bishop's theory that more motivated students and higher quality peers may increase both teacher and student effort (Bishop 2004). For example, Kremer found that in Kenya, a girls' scholarship program also produced gains in exam scores for girls with low pretest scores (who were unlikely to win) and boys (who were ineligible to win), while also having a positive effect on teacher attendance (Kremer, Miguel, and Thornton 2004).

Incentive schemes may not have their intended impact on student learning for several reasons. First, if students lack resources, knowledge or self-control to successfully convert effort into measurable achievement, or also require the support of effective teachers, engaged parents or peer dynamics to learn, then incentives may have little impact (Fryer 2010). The size of the treatment impact may thus differ according to variations in complementary inputs to students' education production function (Fryer 2010). In accordance with this theory, in a large U.S.-based study Fryer tested the impact of offering incentives for educational *inputs,* such as reading a number of books, compared to impact of incentives for educational *outputs,* such as improved test scores, and found that only input-based incentives raised student achievement (Fryer 2010).

Second, external rewards may serve to undermine students' intrinsic motivation to learn, thereby resulting in negative outcomes. Leuven et al. (2010) found evidence of reduced intrinsic motivation in their study of Dutch college students, particularly for lower performing students.

However, a number of other studies have not found any significant negative impact on intrinsic motivation (Fryer 2010; Kremer, Miguel, and Thornton 2004; Sharma 2010; Bettinger 2010).

A third theory holds that academic performance may diminish or worsen once rewards are removed, as incentives may positively effect short-term motivation but negatively influence longer-term motivation (Kohn 1993 cited in Fryer 2010). While no evidence is documented that incentive programs result in decreased long-term motivation, several studies found a fadeout effect once the incentive is removed. For instance, in Kremer's Kenyan scholarship study, academic gains diminished but persisted following the competition (Kremer, Miguel, and Thornton 2004). Fryer (2010) also found evidence of results returning to pre-intervention levels.

Conceptual Framework

This paper offers a conceptual model of student achievement drawing from literature on education in marginalized communities, behavioral economics and past student incentive studies.

Intervention context

Research has shown that intrinsic motivation for learning is largely shaped by family background and socioeconomic status (Sharma 2010; Berry 2011). In addition, students from low socioeconomic backgrounds often exhibit less extrinsic motivation to study, resulting in part from lower perceived returns to education and high discounting of the future (Nguyen 2008). Given the context of rural Chiapas, we anticipate that *telesecundaria* students likely fit Fryer's description of individuals for whom incentives are particularly effective: those with low intrinsic motivation to learn, a lower likelihood of future thinking, and low perceived returns to education (Fryer 2010). Following Fryer (2010) and Bishop (2004), we propose that achievement-based incentives will improve performance by increasing extrinsic motivation to study.

Incentive design

As seen in the literature on incentive eligibility, we expect the program design to greatly influence which students are incentivized to exert additional effort. As incentive schemes that only reward high-performing students run the risk of disincentivizing low performers, the Star Program aims to motivate both high-performing students (through a raffle for students with grades above a certain threshold) and students across the performance spectrum (through an award for most-improved student). The use of short-term rewards, in particular, is designed for students who highly discount the future or have pessimistic perceptions of future opportunities.

Students' grade level may lead to differences in outcomes as younger students may have less discipline for a long-term goal and may be less able to translate their effort into results. This difference may also reflect the impact of different types of incentives on student achievement, as first and second year students are provided with non-monetary incentives, while third year students receive cash awards. Providing awards directly to the student, rather than the parent, is designed to overcome issues regarding child versus parent control (Berry, 2004). Finally, the bulk of the scholarship is provided only once students confirm their enrollment in high school. Conditioning the monetary reward on students' school continuation is designed to reduce the chance that parents will use the funds for their own household expenses.

Channels of influence

As noted by Kremer et al. (2004), the direct behavioral change likely to result from a merit award is increased study effort on the part of the student. Yet, recognizing that student achievement is not determined by effort alone, but by additional factors including the presence of high quality teachers, an engaging curriculum, parent attention, and the behavior of other students in the class, we expect that student achievement will be indirectly impacted by

complementary inputs in the form of effort exerted by parents, teachers and classmates (Fryer 2010; Kremer, Miguel, and Thornton 2004). The mechanisms through which incentive programs might impact these additional channels of influence are described below.

Parent involvement. Kremer et al. (2004) note that as parents become aware that their family can benefit from the incentive program, they may become more attentive to their child's studying and/or more involved in their child's schooling, for example by encouraging their child to complete his/her homework or by attending parent-teacher meetings.

Teacher effort. If incentives motivate students to work harder in their studies, this may make the teaching experience more enjoyable and thereby encourage teachers to increase their effort as well (Kremer, Miguel, and Thornton 2004). It is also possible that as parents become more aware of the possibility that their child could win an award, they will place more pressure on teachers to follow through on their teaching responsibilities, an occurrence that Kremer et al. documented in Kenya. Additionally, if teachers experience benefits from having winners in their class, such as social prestige or gifts from parents, this may also spur them to increase their effort (Kremer, Miguel, and Thornton 2004). If teacher attendance and teaching quality improves, it follows that all students in treatment schools, including those ineligible or unlikely to win awards, stand to benefit from the program (Kremer, Miguel, and Thornton 2004). Kremer finds significant impacts of treatment on teacher attendance and effort, which he proposes may partly explain the positive effects of the program on boys and lower-performing girls. This is consistent with Bishop's model of student learning, wherein school quality and student effort interact positively (Bishop 2004).

Peer effects. The overall academic performance of students is also likely to be impacted by the study effort of other students in the class, since it may be easier to learn in a classroom

with a better disciplinary climate and where classmates are exerting more effort (Kremer, Miguel, and Thornton 2004; Bishop 2004). In other words, students whose classmates are more studious learn more (Bishop 2004). Therefore, to the extent that the incentive program motivates even a few students to study more, we would expect this effect to carry over to other students in the class who may not be directly motivated by the incentive scheme.

In sum, as illustrated in Figure 1, the conceptual model for the study proposes a direct pathway of influence on student achievement through increased student motivation and effort, and indirect influence on student achievement through enhanced teacher and parent involvement.

3. Program Context and Description

Chiapas in Context

The state of Chiapas lies at the southernmost tip of Mexico, sharing a border with Guatemala. The state's population of 4.8 million has an average per capita GDP of MXN $34,751 (USD $1,050), placing it last amongst Mexican States (National Institute of Statistics and Geography 2012). Nearly half of the working population earns less than the daily minimum wage (approximately MXN $60 or USD $4.40), with 17% of workers earning no income (National Institute of Statistics and Geography 2012; Harrup 2011). Over a quarter of households lack running water, and 17% have no sewage or drainage system.

Nearly half of the population of Chiapas lives in rural areas (National Institute of Statistics and Geography 2012). The state is home to a prominent indigenous population, with 27% of residents speaking an indigenous language. Nearly 40% of the indigenous population cannot read or write. Over 80% of the indigenous population works in agriculture, and the per capita income of indigenous residents is approximately one third of the per capita income of the non-indigenous population (National Institute of Statistics and Geography 2010).

Rural Education in Chiapas

Students in rural Chiapas face distinct socioeconomic and geographic disadvantages, as well as critical shortcomings in the quality of classroom instruction. Literacy, school attendance, and average years of schooling in Chiapas all trail national averages (National Institute of Statistics and Geography 2011). To address the particular challenge of reaching rural students, Mexico utilizes a *telesecundaria* model that mixes formal and distance education modalities, wherein lessons are augmented by pre-recorded television lessons and activities. Each classroom has one full-time teacher for all subjects, as opposed to subject-specific teachers used in traditional secondary schools (*secundarias generales*).

To a large extent, the expansion of *telesecundarias* was responsible for the past decade's rise in education coverage (Rincon-Gallardo 2010). Between 1993 and 2006, approximately half of the increases in secondary schools were *telesecundarias* and one in every five secondary students currently attends a *telesecundaria* (Martinez R. 2005; Rincon-Gallardo 2010). Yet while *telesecundarias* contributed to increases in education coverage, the modality is characterized by lower learning outcomes: across OECD countries, 21% and 19% of students were categorized in the lowest two levels of mathematics and reading comprehension, respectively; within Mexico's *telesecundarias*, 95% and 89% fall into the same categories (Vidal and Díaz 2004; Martinez R. 2005). An important caveat to this discrepancy is that rural students – particularly those in families without previous formal education – often lack the resources and extrinsic motivation that benefit their urban counterparts. The education deficit in rural communities must thus be considered both in terms of in-class instruction methods and students' socioeconomic reality.

Star Program

The Star Program is a performance-based incentives program aimed at increasing reading and mathematics learning outcomes in rural middle school populations in Chiapas, Mexico. The Program began a pilot year in October 2010 as a joint venture between the State Government of Chiapas and Foundation Escalera and included 135 *telesecundaria* schools in rural zones with ratings of "High" or "Very High" marginalization based on literacy, health and sanitation, and economic opportunity (Consejo Nacional de Población 2011). In October 2011 the Star Program was reinstated and expanded to 147 *telesecundarias*, totaling 7,852 students. Table 1 (see Appendix) presents demographic characteristics of our sample, showing that students in come from very rural backgrounds, with 91% of families having land for farming or livestock. Nearly 80% of students reported speaking an indigenous language at home. Only 32% of schoolteachers and 22% of directors reported speaking the school's most commonly spoken language. Although televisions are crucial to the method of distance learning, 13% of schools reported lacking electricity and 40% reported broken or missing televisions.

At the beginning of the academic year, Escalera and Ministry of Education personnel visited each school to train students, teachers, and directors in the rules of the program. Teachers were also given explanatory posters to hang in each classroom.

Winner selection was grade-wide, meaning students competed across class groups in a common pool of candidates. Winners could not win by both methods in the same prize period, but could win in successive prize periods. Students were eligible to win awards in two ways:

> *Most Improved Student* (1 Female, 1 Male): The students in each grade who achieve the largest improvement in grade point average (combined, all subjects) between the previous and current marking period.
>
> *A+ Lottery* (1 Female, 1 Male): Every 9.0 grade or higher (on a 10 point scale) in any individual subject merited a distinct ticket in a grade-wide raffle. Separate lotteries were

held for males and females, and unsuccessful tickets could not "carry over" into the following period's lottery.

Prize periods coincided with the second, third and fourth bimesters in a five-bimester academic year. At the conclusion of each prize period, each treatment school carried out a prize ceremony with Escalera representatives. Each ceremony announced and awarded the most-improved students and selected the lottery winners. In 2011-12, the Star Program redesigned the prize scheme according to polled student preferences from the pilot year. Prize types varied by prize period and grade level, as illustrated in Table 2.

4. Evaluation Design

The evaluation of the Star Program follows a randomized controlled trial design. In the 2010-11 pilot year, a sample size of 135 schools was randomly selected from a complete list of public *telesecundarias* in the state's central, northern, and eastern zones. In 2011, the program added 12 additional schools, randomly selected from the state's central highlands zone. The final 2011-2012 sample was set at 76 treatment and 71 control schools.

Monitoring visits by Escalera began in treatment schools mid-year, between one and two months after the initial program training. Teachers and directors were given a brief quiz on the program's details, and a refresher course if needed. Surveyors also ensured two instructional program posters were hung prominently in each classroom.

In May and June of 2012, Escalera survey teams administered endline exams and surveys to both treatment and control groups. Each student was presented with two exams: a math test and a reading test. Test questions were taken from the Texas Assessment of Knowledge and Skills (TAKS) for U.S. Grade 6, a validated instrument with an original Spanish language version. Escalera modified the phrasing of some questions to make the vocabulary more

accessible to students from Chiapas. School visits occurred at school during school hours. Teachers, directors and students completed separate written surveys in separate rooms. One class per grade received surveys. In case of multiple classes per grade, one class was chosen at random to survey. Written student survey questions were distributed to each student and questions were read aloud by surveyors. Written exam questions were read and answered individually and silently. All surveys were filled out on paper, with students using a multiple-choice Scantron response sheet.

As in any experiment it is possible that the simple knowledge of being part of a study may have influenced survey and test results (Hawthorne Effect). To account for the potential bias of experimentation effects, which we expect may be greater for students and teachers in treatment schools, treatment school surveys and exams were completed before the final prize ceremonies to lessen the possibility that ceremony activities would influence testing.

Because of unexpected teacher trainings, local holidays, paydays, and teacher/director sick days, school visits often had to be rescheduled. In the case of three control schools and one treatment school it was not possible to reschedule before the end of the school year, so these schools were not included in the final analysis. One treatment school also discontinued participation in the program, seemingly because parents misunderstood the program rules and asked to be removed from the study. Despite this attrition, our randomization analysis demonstrated that treatment and control schools remained comparable. It is also important to note that classroom conditions were not always conducive to surveying (i.e. rainwater entering the classroom, loss of electricity), which led to additional challenges for data collection.

Data Description

To estimate student performance, we calculated each student's relative math and reading test scores and constructed a combined test score so that each score took a value between zero and one. We normalized these scores using the mean scores and standard deviations of the control group, a standard approach to enable comparisons of results across studies (Kremer, Miguel, and Thornton 2004).

For our control variables, we used responses from the student, teacher and director surveys and also constructed additional variables using student survey results. For instance, to identify the proportion of indigenous students, we classified students as "indigenous" if they reported speaking any language other than or in addition to Spanish at home. The only external control variable used in the study is the Marginalization Index, a variable constructed by Mexico's National Population Council (CONAPO) that uses a wide range of indicators associated with poverty and socioeconomic isolation at the locality level.

Randomization Checks

Following initial assignment of treatment and control groups, the comparability of the groups was tested on a number of school-level characteristics from the 2009-10 school year collected by the Chiapas and Federal Ministries of Education, and the sample was found to be adequately randomized. Variable for randomization checks included student-teacher ratios, gender balances, and ethnic mix, as well as the Marginalization Index and 2010 national standardized test (ENLACE) scores shown in Table 3.

Two methods were used in order to determine whether the randomization remained successful for the 2012 endline study. First, we used two-tailed t-tests to compare the two groups on variables identified from the literature related to student, teacher and director

demographics; school infrastructure and supplies; and other general school characteristics. Next, following Berry (2011) and Fryer (2010), comparability was also checked using the following linear regression model at the school level *s*:

$$TreatmentSchool_s = \alpha + \delta X_{is} + \varepsilon_{is} \quad (1)$$

where the dependent variable is treatment status at the school level, X_{is} are the demographic and school-level control variables and ε_{is} is the error term. For both methods standard errors were clustered on the school level.

Our analyses demonstrated that overall, the treatment and control groups were well randomized, with no significant differences found for variables hypothesized to be most closely associated with student achievement. To address the possibility that non-random characteristics might bias the treatment effect, our econometric strategy, described below, controls for those variables found to be significantly different between groups at the 10% level as well as variables identified as most relevant for student achievement. A randomization table showing t-tests can be found in Table 3.

5. Econometric Specification

General Treatment Effect

Due to time constraints in program implementation and an effort to reduce administrative costs, the study does not have a baseline examination. As such, we cannot control for unobservable differences in personal characteristics (e.g. student ability) by exploiting a two-period study using an individual fixed effects model or a difference in differences analysis.

Instead, to estimate the average program effect on student learning, we will follow Kremer, Miguel and Thornton (2004) in using a Generalized Least Squares (GLS) regression with random errors clustered at the school level. This regression makes use of multiple student

observations within each grade and school to account for potential imperfect randomization on unobservable school characteristics (such as teacher ability). Because GLS regressions absorb part of the variation in the process, estimates can be considered as conservative (Wooldridge 2001). Our basic specification is given in Equation (2):

$$TestScore_{igs} = \alpha + \beta TSchool_s + \delta X_{igs} + \mu_s + \varepsilon_{igs} \qquad (2)$$

where dependent variable $TestScore_{igs}$ is the normalized test score of student *i* in grade *g* in school *s*. Our primary variable of interest, β, reflects the average treatment effect on test scores. Treatment status is captured by dummy variable $TSchool_s$ which is 1 for treatment and 0 for control schools. α is the constant. To account for the joined influence of unobservable school characteristics on students within the same school, standard errors are clustered at the school level. The error term is thus made up of μ_s, a uniform school level error term and a residual error term ε_{ist}, reflecting unobserved student ability and idiosyncratic shocks (characteristics affecting the entire school) (Kremer, Miguel, and Thornton 2004). Lastly, X_{igs} is a vector of control variables. If randomization was successful, these controls will not impact the estimates substantially, yet including them can improve precision of the estimates (Duflo, Glennerster, and Kremer 2006). As described earlier, controls were divided into three sub-categories:

1) Student level variables, including language dummies and socio-economic conditions;
2) Teacher and director characteristics, such as teacher experience and level of education;
3) School and regional level characteristics, such as school supplies availability.

Subsequently, within each sub-category, we included variables that significantly differed between treatment and control schools as well as randomized variables that are hypothesized to be most closely associated with student achievement.

Treatment effects and student ability

We will also consider treatment impact differences across student ability. Here, we follow Hermann and Horn (2011) in using a quantile regression to analyze explanatory variable effects at different points in the conditional distribution of the dependent variable, given in Equation (2):

$$\left(TestScore_{igs} \middle| q\right) = \alpha^q + \beta^q TSchool_s + \delta^q X_{igs} + \varepsilon_{igs} \qquad (2)$$

Where q denotes the quantile of the test-score at which the model is estimated. Hence, when q=25, the treatment effect is estimated on test scores at the 25th quantile test score and its respective control variables. To account for school-level shocks, results were clustered on school level using the bootstrap method (Chen, Wei, and Parzen 2002).

While the use of quantile-regressions is a common practice in estimating heterogeneous treatment effect across student ability (Sharma 2010; Leuven, Oosterbeek, and Klaauw 2003), they are often estimated using baseline test-scores. Given that we only have a post-treatment observation for each student, we cannot directly interpret our regression as a treatment effect for different quantiles as students who may have been classified as low-performing before treatment could have raised their scores to mid-performing at the endline assessment, for instance. Yet given that the program is designed to benefit both low- and high-performers, we would expect that any increase in scores during the school year to be consistent across all subsamples of students, thus limiting concern for bias.

Channels of Influence

Our theoretical framework notes that the program might also have secondary effects on teachers and parents. We will analyze the impact of treatment status on other variables such as teacher attendance using two methods. First, we use an ordered logit regression with clustered standard errors on school level to estimate shifts in response distribution of ordinal variables. All

regressions control for student-level variables, teacher and director characteristics, and school and regional-level characteristics. To estimate program impacts, we report the estimated mean responses for treatment and control groups and identify their difference as the treatment effect. To demonstrate statistical significance, we will also report the coefficients and their standard errors. Secondly, we use a logit regression with clustered errors to estimate the impacts when collapsing the ordinal scales into dummy variables.

6. Results

General Treatment Effect on Test Scores

An overview of estimated general treatment effects on test scores is presented in Table 4. Here we see that the program is associated with a rise in overall test scores of 0.237 standard deviations for treatment school over control schools, significant at the 5% level. The program shows equally sizeable results in mathematics test scores, which are 0.238 standard deviations higher for treatment schools and significant at the 1% level. For the Spanish reading test we see an increase of 0.182 standard deviations, significant at the 5% level. This suggests that while the program had a sizeable and statistically significant average impact, the impact is larger for math scores than reading scores. Recalling our literature review, this finding is in line with the theory of "strategic effort allocation," wherein students confronted with rewards for good grades may react strategically by playing to their strengths and maximizing their expected gains in the short-run (Sharma 2010). Because students in Chiapas generally speak an indigenous language as their first language, the Spanish reading test may thus be relatively more complicated than mathematics and they may choose to exert less effort to increasing their reading performance.

Results by gender

The larger impact on math compared to reading also persists for our regression estimates when broken down by gender. Consistently, the coefficients of treatment impact are larger for math than reading. For boys, treatment status is associated with a 0.244 standard deviation higher test score in math and 0.213 higher score in reading (significant at the 5% and 10% levels, respectively). For girls, the treatment effect on math scores is 0.233 (5% significance), while the reading treatment effect is 0.154 standard deviations higher, though statistically insignificant.

A further difference shown in Table 4 is that the treatment effect seems larger for male students. For boys, we find a statistically significant rise in overall, math and reading scores at conventional levels. Yet for girls, the treatment coefficients are lower and there is no significant impact on reading scores. This may be explained in part by the theory of complementary inputs, which, as described in conceptual framework, posits that student achievement may require additional inputs such as parental and teacher support (Fryer 2010). If female students receive less support, they may be less able to turn their additional effort into increased test scores.

Results by grade

To further analyze the program's effects on test scores, we broke down the results by grade level (see Table 5). As noted in the conceptual framework, we might expect treatment effect differences across grades both because of differences in maturity and discipline, as well as the different types of awards for first and second grade students (material prizes) compared to third grade students (financial prizes).

Table 5 shows there are indeed large differences in the program's impact on overall test scores across grades. For first and second grade students, the program is associated with only a small impact at 0.162 and 0.0375 standard deviations, respectively, and both are insignificant at

conventional levels. In contrast, for third grade students, the program had a sizeable impact of 0.309 standard deviations, which is statistically significant at the 1% level.

Interestingly, when we break down the results by gender for each grade, we see a somewhat different pattern arising than from our overall analysis. Namely, both for first grade and third grade the treatment effect on overall test scores was larger for girls (0.225 and 0.391, significant at the 5% and 1% level, respectively) than boys (0.171 and 0.200, both significant at the 10% level), while there is no significant treatment effect for boys or girls in second grade.

When comparing the program impact further by subject (mathematics versus reading), we see certain patterns that are driving the higher female averages in Grade 1 and Grade 3. In Grade 1, the only significant impact we find is for girls in mathematics (0.181 standard deviations, significant at the 10% level). Girls in Grade 1 show no significant impact in reading, and boys show no significant treatment effects in either subject.

In contrast, in Grade 3 we observe a sizeable and statistically significant effect for both boys and girls in both mathematics and reading. The main gender difference in Grade 3 seems to be the *size* of the treatment effect, which is consistently bigger for girls (0.334 in math and 0.346 in reading, significant at the 1% and 5% level, respectively) than for boys (0.270 in math and 0.224, also significant at the 1% and 5% level, respectively). The impacts we observe in Grade 3 are closest to those one would expect given a successful incentive program. A potential explanation for this is that the stakes are higher for these students as they are eligible to win up to $2,000 pesos ($150 USD) in scholarships, while winners in Grade 1 and 2 could only win books, MP3-players and mini-laptops. This suggests that financial incentives may be more effective in motivating student learning as compared to material incentives. The impact of financial

incentives on parental support may also partly explain the higher impact seen for Grade 3 girls. This channel of influence will be discussed further in the following section.

The results in Grade 2 offer something of a mystery. Here the only significant overall impact lies with mathematics (0.227, significant at 5% level), which is driven by males (0.421, significant at the 1% level). Boys see no significant impact in reading and there is no significant impact on girls in either math or reading. Following theories on complementary support and the power of financial incentives (Fryer 2010; Bishop 2004), girls in Grade 2 may not receive the same level of support from their parents as they would in the case of financial incentives (which benefit the family), and therefore may not experience treatment effects comparable to Grade 3.

Treatment effects and student ability

Beyond identifying the overall treatment impact on test scores, the success of the Star program also depends on its ability to influence student learning for the whole classroom. As noted earlier, several studies found that incentive programs benefitted high-performing students while disincentivizing lower performers (Leuven, Oosterbeek, and Klaauw 2003). As discussed previously, to avoid such an outcome the Star Program was explicitly designed to incentivize both high-performing and low-performing students. To assess whether this combination of rewards had its intended effect, we divided the sample into five "quantiles" of students' relative test performance: the lowest 10%, 25%, the median performers, and students with the highest 75% and 90% of scores. We then estimated the treatment effect on test performance for students with different test scores, shown in Table 6.

Panel A of Table 6 presents the joint estimates for all grades in five quantiles for overall test scores, math and reading. Only the lowest performing 10% of students in each subject sees a statistically significant rise in estimates for all three scores (0.171 standard deviations in overall

scores, 0.244 standard deviations in mathematics and 0.194 standard deviations in reading). For the lowest performing 25% of students, we only observe a significant effect for mathematics (0.320 standard deviations). The median student sees both a significant overall impact and an impact in math scores (0.278 and 0.292 standard deviations), while the highest performing 75% and 90% only see a significant increase in math (both 0.244 standard deviations).

In panels B, C and D, we see the estimated quantile-regressions for Grades 1, 2 and 3. Here we observe that our main findings for general treatment effects are also present for these estimates. In other words, the program generally has a higher impact on mathematics than reading scores, and the program is substantially more effective for third grade students. Hence, while the impacts on overall scores are generally insignificant for first and second graders, third grade students from treatment schools with the lowest 10% and 25% performances see a 0.338 and 0.296 standard deviation rise in overall scores, median treatment students see a 0.319 standard deviation rise, and we observe a 0.356 and 0.274 rise in overall test scores for the highest 75% and 90% performers.

The results broken down by quantile provide us with two important conclusions. Firstly, when the program has a substantial impact (i.e., Grade 3), we see an overall treatment effect for the whole classroom. This suggests that the mix of incentives offered by the program is successful in motivating both low- and high-performing students. Secondly, considering solely the size of the coefficients for math and reading (Figures 2, 3), we see that treatment effects for mathematics are higher for lower-performing students than high-performing students, who thus appear to focus their effort more on mathematics. In contrast, for reading we see higher coefficients for high-performing students, suggesting they allocate more effort towards reading.

It is also important to note that significance levels for these results are very conservative, as each estimate runs a different regression with only a small sub-section of observations. To account for school-level shocks, these results were also clustered for each quantile on school level using the bootstrap method (Chen, Wei, and Parzen 2002). Because this combination of quantile regressions and bootstrapping significantly reduces the degrees of freedom for the estimations, it is unsurprising that a large share of our estimates is statistically insignificant.

Channels of Influence

Our analysis suggests that the Star program had a sizeable and statistically significant impact on raising students' overall test scores. Throughout this section, we will try to identify some of the mechanisms through which the program influenced student performance. Recalling our conceptual framework, we noted that student learning required both student effort and certain compliments, and laid out several mechanisms for identifying the importance of such *indirect* channels in explaining the program's impact, we analyze program effects on variables related to teacher effort, parental support and student motivation.

Indirect treatment effects through teachers

As our conceptual framework describes, an award scheme might improve teacher effort through informal social sanctions and/or rewards from parents, as well as by through the positive effects of teaching more motivated students. In Table 7, we first show the treatment effects on student-reported figures of teacher attendance (student-reported figures were considered to be less biased than teacher-reported figures). By comparing the difference in the estimated mean of student responses for treatment and control schools, we can analyze the program impact on teacher attendance. We see that students in treatment schools are 13.44% more probable than control schools to have reported no teacher absence in the last four weeks, and 14.68% more

probable to have reported teacher absence of 3 days or fewer, with estimated coefficients that are significant at the 1% level.

To further explore whether teachers in treatment schools also exerted greater effort in their classroom, we used a simple proxy of how often teachers asked students if they understood the topic presented to them, as reported by students. Here we find that students in treatment schools had a 5.43% higher probability of reporting "always" and a 10.65% higher probability of reporting "always or almost always" being asked if they understand their topics, both significant at the 10% level. Hence, we find substantial evidence that the program affects the level of teacher attendance and teacher effort, which may provide an indirect channel through which the program raises test scores.

One potential limitation of these estimates is that due to the lack of a baseline we have no information on teacher attendance or effort before the program, and thus cannot safely attribute this difference to the program alone. However, an initial assessment carried out in the pilot year suggested that teacher attendance was comparable for treatment and control schools.

Indirect treatment effects through parents

As noted earlier, the Star Program may incentivize parents to improve student performance if they become aware that their family can benefit from the program, and therefore become more attentive to their child's schooling. To identify whether our program offers support for these hypotheses, we first analyzed the program's impact on the question, "My parents have attended a school meeting this year." In Table 7, Panel B, we see treatment schools have a 5.57% higher estimated probability of students reporting "yes" to this question, with a coefficient that is significant at the 5% level.

This proximate variable for parent involvement is also important to test two hypotheses on gender bias. When we break down the responses by gender in the control group, we see that parents of girls are substantially less likely to have attended school meeting than boys (76.77% versus 83.4%). This supports our theory on treatment effects across gender, as boys commonly receive more parental support than girls with respect to schooling. Secondly, when we consider differences in treatment effect on parental support for boys and girls, the effect is considerably larger for girls (an 11.1% increase in probability of reporting "yes") compared to boys (a 4.7% increase). By producing a higher increase in parental support for girls compared to boys, the program may thus be offering a counterbalance to gender bias in parental support.

A third hypothesis related to parental involvement centers around material and financial incentives. As noted in Berry (2011), parents are more likely to appropriate children's scholarships than in-kind prizes. In this way, parents of Grade 3 students may have a greater incentive to be involved in their child's education if the family benefits more from monetary prizes. Comparing the treatment impact of 2.85% on parents attending school meetings for Grade 1 and 2 (receiving material incentives) with the treatment impact of 6.26% for Grade 3 (financial incentives), we find evidence suggesting that parents care more about the latter. A limiting factor in these tests, likely resulting from reduced sample size, is that when we break down parental involvement by gender or grade, the estimates become statistically insignificant.

Another mechanism through which parents might influence test scores is by checking up on their children (Kremer, Miguel, and Thornton 2004). Hence, another indicator of parental involvement is the question, "My parents keep informed of my grades in school." In Table 7, Panel B, we see that treatment is associated with a 9.3% higher response of children saying "yes", however the finding is statistically insignificant. We also see a similar difference in

treatment effect across gender, with an estimated 7.97% increase for boys saying "yes" compared to 11.10% for girls. There are no observable differences in coefficients for Grade 1 and 2 averages versus Grade 3.

Additional treatment effects for students

In the previous sections we saw that the Star Program is associated not only with a sizeable increase in test scores but also with an increase in teacher effort and parental support. Indeed, though we cannot be certain, perhaps the reason we see such rises in test scores is because of these indirect channels of influence. While these results are important in and of themselves, they are even more interesting when considering that we see very little differences in student motivation between treatment and control groups. To measure student motivation, we will consider four different variables. We begin by analyzing student attendance, followed by student comprehension of classes taught, and lastly we will report on students having paid work outside of school as well as their hours of paid work (results shown in Table 8).

When considering student attendance, we used both student- and teacher-reported answers on the number of days students were absent in the last four weeks. Yet, for both, results are almost identical for treatment and control groups and any treatment effect is statistically insignificant. The program thus appears not to have any impact on student attendance.

Regarding student comprehension, we find that students in treatment schools have an 8% higher probability of saying they "always or almost always understand their classes," yet this effect is not significant. The coefficients are similar in size when broken down by gender.

Lastly, the program may make students want to reduce their paid work outside of school so they have more time studying. Hence, if paid work is voluntary (a questionable assumption), it provides an inverse proximate variable for student effort. If it is family-induced, a reduction in

paid work could be better interpreted as family support for the child's learning. In Table 8 we show that students in treatment schools are 6.74% less likely to work, at 5% significance. They also appear to work fewer hours than control school students, yet this estimate is insignificant.

There are several limitations to keep in mind when measuring student motivation. First, responses are student-reported and therefore subject to bias. For instance, students might fear that they will get into trouble if they report their actual level of attendance, and thus provide artificially high attendance rates. Similarly, for our variable on class comprehension, 80.3% of the control school students and 88.3% of the treatment school students responded that they always or almost always understand their classes, even though our test average would suggest otherwise. A third bias is present for the question on paid work, as it is often unclear if assisting the family classifies as paid work.

7. Conclusion

In recent years several studies have examined the effectiveness of providing incentives based on student achievement in the classroom. This paper reports the results of a mixed-incentive program in rural middle schools in the Mexican state of Chiapas. We find that the Star Program had a sizeable, significant and robust impact on student performance, raising overall test scores by 0.237 standard deviations. The program was particularly influential in improving math scores, producing a 0.238 standard deviation increase in test scores for treatment schools, and an increase of 0.182 standard deviations in the Spanish reading test. We also found that in cases where the program has a substantial impact (such as on math scores), we see that there is a significant treatment effect for low, average and high-performing students. This suggests that the program's mix of incentives is successful in motivating the whole classroom. Finally, we observe large differences in effect size across grades, which may reflect different motivational

impacts for material and financial prizes. The program is associated with only a small and insignificant impact on overall test scores for first and second grade students, while it has a sizeable impact for third grade students. The program thus suggests that students respond particularly well to financial incentives.

Moreover, the study found significant evidence that the treatment had an impact on teacher effort and parental support. Firstly, students in treatment schools were nearly 15% more likely to have reported lower teacher absence (three days or fewer) in the last four weeks. Secondly, teachers also appeared to exert more effort in treatment schools, with students being more than 10% more likely to report that they were "always or almost always" asked if they understood their topic. Thirdly, students in treatment schools were nearly 6% more likely to report that their parents attended a school meeting in the last year.

This study thus follows Kremer et al. (2004) and Fryer (2010) in holding that student achievement is not determined by effort alone, but rather requires the complementary support of high quality teachers and engaged parents to ensure learning. We believe that the impact of the Star Program on teacher effort and parental support highlight important *indirect* channels of student learning that help to explain the program's impact on test scores. In line with recent studies on educational performance (Glewwe, Kremer, and Moulin 2007; Kremer, Miguel, and Thornton 2004; Friedman et al. 2011), these results suggest that we cannot identify the impact of educational programs without considering their complementarities with, and impact on, wider community involvement in schools.

A key area for future study of the Star Program would be to replicate our findings in subsequent years, while improving the precision in identifying our channels of influence. The follow-up program has been designed to have a baseline measure, which will allow us to further

validate impacts across the classroom, as well as influences on teacher effort, teacher attendance, and parental involvement. A future study might also experiment with offering financial incentives in all grades, to estimate whether differences in impact across grades were found because of differences in the incentives offered or because of age-specific reasons. Subsequent research might also include a complementary qualitative study, which would aid in evaluating program effects that are difficult to measure quantitatively, such as the impact on students' intrinsic motivation, possible peer effects, mechanisms through which the program impacts teachers and parents, and how these indirect impacts in turn influence student learning.

Finally, given that our study's sample seems adequately randomized, it provided a robust analysis with appropriate internal validity. However, because the study chose the most marginalized regions of Chiapas, which itself is a very poor state, the sample is not representative of the wider Mexican context. To identify whether financial incentives are equally successful in less marginalized communities, it would be worthwhile to initiate similar incentive programs in other Mexican regions.

In sum, incentive programs offer promising results for improving student achievement in the classroom and provide a fruitful basis for future research. By better understanding the various channels through which students achieve higher test scores, be it through direct effects, peer effects or indirect channels of greater teacher and parental support, we can enhance the design of incentive programs and provide even more effective and innovative ways to improve student learning.

References

Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review* 99 (4) (August): 1384–1414. doi:10.1257/aer.99.4.1384.

Behrman, Jere R., Susan W. Parker, and Petra E. Todd. 2005. *Long-Term Impacts of the Oportunidades Conditional Cash Transfer Program on Rural Youth in Mexico*. Ibero America Institute for Econ. Research (IAI) Discussion Paper. Ibero-America Institute for Economic Research. http://ideas.repec.org/p/got/iaidps/122.html.

Berry, James. 2011. "Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India". Unpublished manuscript. Cornell University.

Bettinger, Eric P. 2010. *Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores*. Working Paper. National Bureau of Economic Research. http://www.nber.org/papers/w16333.

Bishop, John. 2004. "Drinking from the Fountain of Knowledge: Student Incentive to Study and Learn-Externalities, Information Problems and Peer Pressure." *CAHRS Working Paper Series* (October 1). http://digitalcommons.ilr.cornell.edu/cahrswp/19.

Chen, Li, Lee-Jen Wei, and Michael I. Parzen. 2002. *Quantile Regression for Correlated Observations*. Technical Report. University of Minnesota.

Consejo Nacional de Población. 2011. *Índice De Marginación Por Entidad Federativa y Municipio 2010*. http://www.conapo.gob.mx/es/CONAPO/Indices_de_Marginacion_2010_por_entidad_federativa_y_municipio.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2006. *Using Randomization in Development Economics Research: A Toolkit*. Abdul Latif Jameel Poverty Action Lab.

Friedman, Willa, Michael Kremer, Edward Miguel, and Rebecca Thornton. 2011. *Education as Liberation?* Abdul Latif Jameel Poverty Action Lab.

Fryer, Roland G. 2010. *Financial Incentives and Student Achievement: Evidence from Randomized Trials*. Working Paper. National Bureau of Economic Research. http://www.nber.org/papers/w15898.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2007. *Many Children Left Behind? Textbooks and Test Scores in Kenya*. Working Paper. National Bureau of Economic Research. http://www.nber.org/papers/w13300.

Haisley, Emily, Romel Mostafa, and George Loewenstein. 2008. "Myopic Risk-seeking: The Impact of Narrow Decision Bracketing on Lottery Play." *Journal of Risk and Uncertainty* 37 (1): 57–75. doi:10.1007/s11166-008-9041-1.

Harrup, Anthony. 2011. "Mexico's 2012 Minimum Wage Increase Set at 4.2%." *Wall Street Journal*, December 10. http://online.wsj.com/article/SB10001424052970203413304577091050495760194.html.

Hermann, Zoltan, and Daniel Horn. 2011. *How Inequality of Opportunity and Mean Student Performance Are Related? A Quantile Regression Approach Using PISA Data*. Institute of Economics, Hungarian Academy of Sciences.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2004. *Incentives to Learn*. Working Paper. National Bureau of Economic Research. http://www.nber.org/papers/w10971.

Leuven, Edwin, Hessel Oosterbeek, and Bas Van der Klaauw. 2003. "The Effect of Financial Rewards on Students' Achievements: Evidence from a Randomized Experiment." *SSRN eLibrary* (June). http://papers.ssrn.com.ezproxy.cul.columbia.edu/sol3/papers.cfm?abstract_id=428101.

Martinez R., Felipe. 2005. *La Telesecundaria Mexicana. Desarrollo y Problemática Actual.* Colección Cuadernos De Investigación. Instituto Nacional para la Evaluación de la Educación.

National Institute of Statistics and Geography. 2010. *Statistical Perspectives Chiapas*.

———. 2011. *Statistical Perspectives Chiapas*.

———. 2012. *Statistical Perspectives Chiapas*. http://www.inegi.org.mx/prod_serv/contenidos/espanol/bvinegi/productos/integracion/estd_perspect/chis/Pers-chs.pdf.

Nguyen, Trang. 2008. *Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar*. Massachusetts Institute of Technology.

Ponce, Juan, and Arjun S. Bedi. 2010. "The Impact of a Cash Transfer Program on Cognitive Achievement: The Bono De Desarrollo Humano of Ecuador." *Economics of Education Review* 29 (1) (February): 116–125. doi:10.1016/j.econedurev.2009.07.005.

Rincon-Gallardo, Santiago. 2010. "Some Barriers to Quality of Educational Opportunity in Mexican Telesecundarias". Cambridge, MA: Harvard Graduate School of Education, International Education Policy Program.

Schultz, Paul T. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74 (1) (June): 199–250. doi:10.1016/j.jdeveco.2003.12.009.

Sharma, Dhiraj. 2010. "The Impact of Financial Incentives on Academic Achievement and Household Behavior: Evidence from a Randomized Trial in Nepal." *SSRN eLibrary* (November 21). http://papers.ssrn.com.ezproxy.cul.columbia.edu/sol3/papers.cfm?abstract_id=1681186.

Vidal, Rafael, and M.A. Díaz. 2004. *Resultados De Las Pruebas PISA 2000 y 2003 En México*. Instituto Nacional para la Evaluación de la Educación.

Wooldridge, Jeffrey M. 2001. *Econometric Analysis of Cross Section and Panel Data*. 1st ed. The MIT Press.

Annex: Figures and Tables

Figure 1: Conceptual Framework

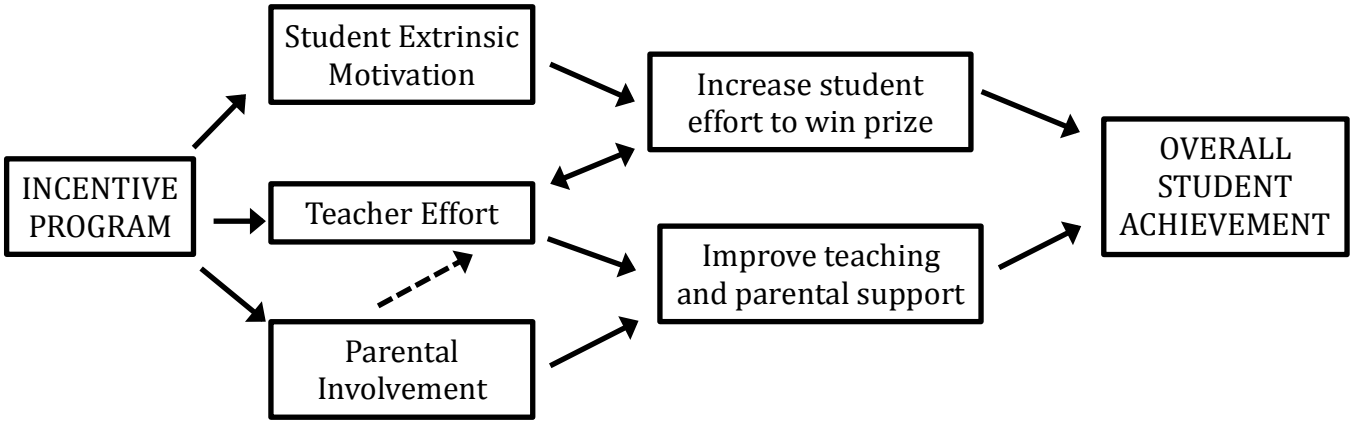

Figure 2: Estimates of Treatment Impact by Achievement Level for Mathematics

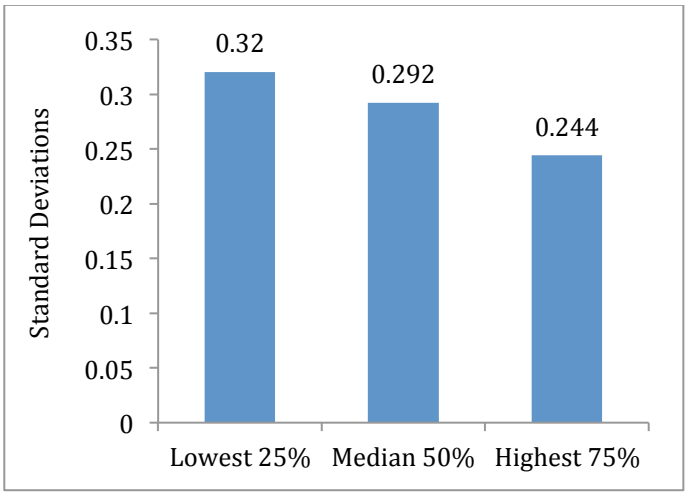

Figure 3: Estimates of Treatment Impact by Achievement Level for Reading

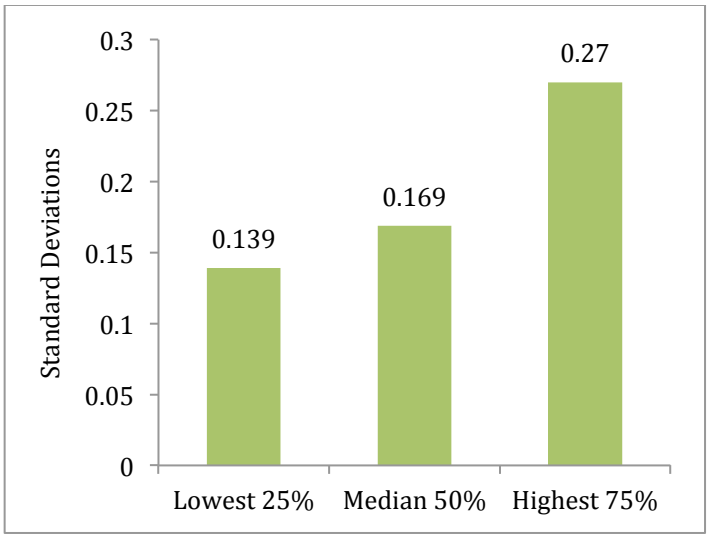

Table 1: Sample Description

Panel A: Demographic Overview	Total	Treatment	Control
Number of schools	147	76	71
Number of students	7852	4011	3841
Percentage female	47.6	47.7	47.4
Percentage indigenous students	79.0	79.7	78.3

Panel B: Population and Environment	Total
Students	
Travel 30 minutes or more to school	24.4%
Travel 1 hour or more to school	11.8%
Participate in some type of paid work	33.1%
Parents with high school education or more*	11.0%
Student families with land for farming or livestock	91.0%
Speak Spanish at home	50.4%
Running water in house	24.9%
Electricity in house	92.1%
Teachers	
Reside full-time in town/village where they work	6.3%
Speaks school's most spoken language	31.6%
Bachelor's degree or higher	47.2%
Directors	
Reside full-time in town/village where they work	9.4%
Speaks school's most spoken language	22.2%
Schools	
Classrooms without electricity	12.8%
Classrooms with broken or missing televisions*	40.0%
Classes without access to computers	29.3%

Notes: * These variables use 2011 survey data because this question was missing from the 2012 survey data. For further detail of the sample description, please see the Randomization Overview (Table 3).

Table 2: Star Program Selection Method and Prize Type

		Prize Period 1	*Prize Period 2*	*Prize Period 3*
	Selection Method	**Prize Type**	**Prize Type**	**Prize Type**
1st Grade	**Most Improved**	Certificate	Certificate + Book	Certificate + 2 Books
	Lottery	Certificate	Certificate + Book	Certificate + 2 Books
2nd Grade	**Most Improved**	Certificate	Certificate + MP3 Player	Certificate + MP3 Player
	Lottery	Certificate	Certificate + MP3 Player	Certificate + Laptop
3rd Grade	**Most Improved**	Certificate	Certificate + 500 Peso Scholarship	Certificate + 2000 Peso Scholarship*
	Lottery	Certificate	Certificate + 500 Peso Scholarship	Certificate + 2000 Peso Scholarship*

Notes: *Scholarships divided into three installments, with the first given at ceremony and the second and third given upon proof of high school enrollment.

Table 3: Randomization Table of Student variables (Panel A)
Teacher and Director characteristics (Panel B)
and School and Regional level variables (Panel C)

	Treatment	Control	Difference
Panel A: Student Characteristics			
Percentage female	47.7%	47.4%	0.003
Percentage indigenous students	79.7%	78.3%	0.014
Percentage Ch'ol spoken at home	40.5%	38.9%	0.016
Percentage Tsotsil spoken at home	4.8%	5.2%	-0.004
Percentage Tzetal spoken at home	34.1%	31.5%	0.026
Percentage Spanish spoken at home	49.3%	51.4%	-0.021
Percentage other language spoken	7.0%	8.1%	-0.011
Percentage grade repetition	13.4%	15.7%	-0.023
Students living with father	76.0%	77.0%	-0.01
Students living with mother	66.1%	65.3%	0.0080
Students living with grandparents	27.1%	28.9%	-0.018
Father has paid employment	92.5%	93.2%	-0.007
Mother has paid employment	75.4%	80.9%	-0.055***
Family receives remittances	35.6%	39.6%	-0.04
Family cultivates land or grows animals	87.7%	86.5%	0.012
Students eating breakfast every day	74.7%	77.1%	-0.024
Panel B: Teacher and Director Characteristics			
Percentage female teachers	50.5%	42.9%	0.076
Average age teacher	30	30	0
Teacher education - Secondary school or lower	54.9%	50.5%	0.044
Teacher education - University or higher	45.1%	49.5%	-0.044
Teacher experience - less than 5 years	49.8%	58.9%	-0.091
Teacher experience - more than 5 years	50.1%	41.6%	0.091
Percentage teachers living in the school's community	5.8%	7.5%	-0.017
Percentage teachers speaking school's most spoken language	27.8%	32.3%	-0.045
Percentage female directors	23.1%	25.5%	-0.024
Average age directors	34	34	0
Director education - Secondary school or lower	59.4%	62.9%	-0.035
Director education - University or higher	40.7%	37.1%	0.036
Director experience - less than one year	46.2%	37.8%	0.084
Director experience - more than one year	53.8%	62.2%	-0.084
Percentage directors living in the school's community	4.8%	11.8%	-0.07
Percentage directors speaking school's most spoken language	18.5%	25.5%	-0.07
Panel C: School and Regional Level Characteristics			
Log Students Total 2010 ENLACE Score	6.9198	6.9043	0.0155
Log Students Total 2010 ENLACE Score Spanish	6.1551	6.1543	0.0008
Log Students Total 2010 ENLACE Score Math	6.2917	6.2637	0.028
Marginalization Index	17.47	18.01	-0.540
Number of students per school	118	113	5
Percentage school with electricity	93.8%	89.1%	0.0469
Percentage schools conntected with piped water	66.7%	60.0%	0.0670
Percentage schools where every student has pencils and pens	92.5%	80.7%	0.118*
Percentage schools where every student has notebooks	93.1%	79.8%	0.133*
Percentage schools where every student has textbooks	85.9%	75.5%	0.104
Percentage schools where every student has a desk	90.8%	87.1%	0.366
Percentage schools where every student has a uniform	74.6%	67.2%	0.074
Number of students per grade	37	38	-1
Number of students per class	23	24	-1
Percentage classrooms with electricity	88.1%	85.4%	0.027
Classroom conditions are good or acceptable	68.8%	74.9%	-0.061
Classroom conditions are bad or very bad	31.1%	25.1%	0.06

Notes: Difference between group means is estimated by a two-sided t-test. Standard errors used in testing difference between group means are clustered at the school level.***Significant at 1%, **Significant at 5%, *Significant at 10%

Table 4: Program Impact on Test Scores

	Combined	Math	Spanish
Overall	0.237**	0.238***	0.182**
	(0.0945)	(0.0851)	(0.0907)
Obs.	7624	7624	7624
Boys	0.259**	0.244**	0.213*
	(0.109)	(0.0975)	(0.110)
Obs.	4007	4007	4007
Girls	0.217**	0.233**	0.154
	(0.103)	(0.0924)	(0.106)
Obs.	3605	3605	3605

Notes: Dependent variables are normalized scores (with respect to mean and standard deviation of test scores of control group). Standard errors are clustered at the school level and reported in parentheses. All regressions control for student level variables (e.g. language dummies); teacher and director characteristics (e.g. years of teacher experience) and school and regional level characteristics (e.g. classroom conditions).***Significant at 1%, **Significant at 5%, *Significant at 10%.

Table 5: Estimate of differential treatment impact by grade

Panel A: Grade 1	Combined	Math	Spanish
Overall	0.162	0.1000	0.109
	(0.110)	(0.110)	(0.0975)
Obs.	2643	2643	2643
Male	0.171*	0.146	0.126
	(0.0895)	(0.119)	(0.0769)
Obs.	1409	1409	1409
Female	0.225**	0.181*	0.157
	(0.113)	(0.106)	(0.0996)
Obs.	1228	1228	1228
Panel A: Grade 2	Combined	Math	Spanish
Overall	0.0375	0.227**	-0.0479
	(0.103)	(0.0887)	(0.103)
Obs.	2524	2524	2524
Male	0.222	0.421***	0.133
	(0.144)	(0.117)	(0.146)
Obs.	1290	1290	1290
Female	0.0179	0.177	-0.0665
	(0.137)	(0.123)	(0.135)
Obs.	1230	1230	1230
Panel A: Grade 3	Combined	Math	Spanish
Overall	0.309***	0.246**	0.301***
	(0.111)	(0.0978)	(0.111)
Obs.	2457	2457	2457
Male	0.200*	0.270***	0.224**
	(0.105)	(0.0879)	(0.0960)
Obs.	1308	1308	1308
Female	0.391***	0.334***	0.346**
	(0.151)	(0.116)	(0.153)
Obs.	1147	1147	1147

Notes: Dependent variables are normalized scores (with respect to mean and standard deviation of test scores of control group). Standard errors are clustered at the school level and reported in parentheses. All regressions control for student level variables (e.g. language dummies); teacher and director characteristics (e.g. years of teacher experience) and school and regional level characteristics (e.g. classroom conditions).***Significant at 1%, **Significant at 5%, *Significant at 10%.

Table 6: Quantile Regression Estimate of Treatment Impact by Achievement Level

Panel A: Overall	Combined	Math	Spanish
Quantile 0.1	0.171**	0.244**	0.194*
	(0.0749)	(0.103)	(0.107)
Quantile 0.25	0.229	0.320**	0.139
	(0.141)	(0.139)	(0.111)
Quantile 0.5	0.278*	0.292**	0.169
	(0.168)	(0.149)	(0.156)
Quantile 0.75	0.284	0.244*	0.270
	(0.191)	(0.130)	(0.167)
Quantile 0.9	0.166	0.244*	0.210
	(0.172)	(0.135)	(0.333)
Obs.	7624	7624	7624
Panel B: Grade 1	Combined	Math	Spanish
Quantile 0.1	0.142*	0.100	0.181
	(0.0861)	(0.201)	(0.147)
Quantile 0.25	0.169	0.163	0.0638
	(0.107)	(0.193)	(0.125)
Quantile 0.5	0.189**	0.146	0.0954
	(0.0958)	(0.218)	(0.230)
Quantile 0.75	0.233*	0.282	0.259
	(0.136)	(0.188)	(0.227)
Quantile 0.9	0.190	0.232	0.253
	(0.165)	(0.250)	(0.288)
Obs.	2643	2643	2643
Panel C: Grade 2	Combined	Math	Spanish
Quantile 0.1	0.157*	0.250	0.0326
	(0.0880)	(0.179)	(0.181)
Quantile 0.25	0.193	0.0982	-0.0136
	(0.261)	(0.193)	(0.210)
Quantile 0.5	0.237*	0.0635	0.0503
	(0.1249)	(0.168)	(0.200)
Quantile 0.75	0.227	0.0802	-0.0717
	(0.270)	(0.172)	(0.175)
Quantile 0.9	0.216*	0.180	-0.0192
	(0.128)	(0.290)	(0.213)
Obs.	2524	2524	2524
Panel D: Grade 3	Combined	Math	Spanish
Quantile 0.1	0.338***	0.429**	0.192
	(0.108)	(0.202)	(0.169)
Quantile 0.25	0.296**	0.344	0.214
	(0.120)	(0.227)	(0.171)
Quantile 0.5	0.319***	0.307	0.266
	(0.102)	(0.211)	(0.205)
Quantile 0.75	0.356***	0.228	0.318
	(0.118)	(0.205)	(0.239)
Quantile 0.9	0.274**	0.210	0.405**
	(0.126)	(0.235)	(0.206)
Obs.	2457	2457	2457

Notes: Dependent variables are normalized scores (with respect to mean and standard deviation of test scores of control group). Calculated using a quantile regression along different points of the conditional test score distribution. Standard errors are bootstrapped and clustered at the school level and reported in parentheses. All regressions control for student level variables (e.g. language dummies); teacher and director characteristics (e.g. years of teacher experience) and school and regional level characteristics (e.g. classroom conditions).***Significant at 1%, **Significant at 5%, *Significant at 10%.

Table 7: Regression Estimates of Treatment Impact on Teacher effort (Panel A) and Parental Support (Panel B)

	Coefficient	Est. mean Treatment	Est. mean Control	Difference
Dependent Variables:				
Panel A: Teacher attendance and effort				
Number days of teacher absence last 4 weeks	-.5694*** (.2055)			
• Zero days		0.6775	0.5432	0.1344
• Between 1-3 days		0.224	0.295	-0.071
• Between 3-5 days		0.0451	0.0712	-0.0261
• Between 5-10 days		0.0455	0.0768	-0.0313
• More than 10 days		0.0079	0.0138	-0.006
Number days of teacher absence last 4 weeks (1=3 days or less, 0=3 days or more)	.7306*** (.2199)	0.9033	0.7564	0.1468
Teacher asks students if they understand topic	-.2371* (.1460)			
• Always		0.671	0.6167	0.0543
• Almost always		0.1857	0.2084	-0.0227
• Almost never		0.0796	0.0956	-0.016
• Never		0.0636	0.0793	-0.0157
Teacher asks students if they understand topic (1=always or almost always, 0=almost never or never)	.3477* (.1926)	0.8896	0.7831	0.1065
Panel B: Parental support				
Parents have attended school meeting this year (yes, no)	.4067** (.1910)	.9352	.8795	.0557
• Male responses	.3330 (.2469)	.9324	.8340	.0470
• Female responses	.3271 (.2146)	.8788	.7677	.1110
• Grade 1 & 2 only	.2327 (.2447)	.9206	.8920	.0285
• Grade 3 only	.17185 (.3915)	.8057	.7432	.0626
Parents keep track of student grades (yes, no)	.1421 (.1979)	.8530	.7560	.0930
• Male responses	.0403478 (.2226)	.8340	.7543	.0797
• Female responses	.327091 (.2146)	.8788	.7677	.1110
• Grade 1 & 2 only	.1674 (.2276)	.8253	.7658	.0595
• Grade 3 only	.1719 (.3915)	.8058	.7432	.0626

Notes: The probability mean for treatment and control schools are estimated using an ordered logit regression. The treatment effect 'difference' is identified by the marginal change in probability when the dummy variable for treatment changes from 0 to 1. All variables in this table come from student-reported data. Mean probabilities for dummy variables are estimated using a logit regression. All regressions are estimated with clustered standard errors at the school level and are reported in parentheses. All regressions also control for student level variables (e.g. language dummies); teacher and director characteristics (e.g. years of teacher experience) and school and regional level characteristics (e.g. classroom conditions). ***Significant at 1%, **Significant at 5%, *Significant at 10%.

Table 8: Regression Estimates of Treatment Impact on Student effort

	Coefficient	Est. mean Treatment	Est. mean Control	Difference
Dependent Variables:				
Number days of student absence last 4 weeks	.0634083 (.1367)			
• Zero days		0.7497	0.7497	-0.0117
• Between 1-3 days		0.1782	0.1782	0.0076
• Between 3-5 days		0.0402	0.0402	0.0022
• Between 5-10 days		0.0225	0.0225	0.0013
• More than 10 days		0.0095	0.0095	0.0006
Level of attendance rate for a teacher's class	.0068452 (.3609)			
• Less than 20%		0.0108	0.0109	-0.0001
• Between 20-40%		0.005	0.005	0
• Between 40-60%		0.0119	0.012	-0.0001
• Between 60-80%		0.1292	0.1299	-0.0007
• More than 80%		0.8417	0.8408	0.0009
Student understand their classes (1=Always or almost always, 0=almost never or never)	.2193685 (.2172)	.883448	.8037298	.0797182
• Male responses	.1828537 (.2459)	.88084	.8068768	.0739583
• Female responses	.2473648 (.2412)	.8876965	.7952411	.0925
Students have paid work outside of school (Yes, no)	-.2886** (.1403)	.28478	.3521383	-.06735
• Male responses	-.29664* (.1550)	.3623	.4411	-.0788
• Female responses	-.3084* (.1655)	.19692	.25932	-.0624
Hours of student paid work students outside of school last week (1=0-3 hours, 0=3 hours or more)	.1976 (.1334)	.6958	.6397	.05613
• Male responses	.1616 (.1771)	.5953	.5296	.0657
• Female responses	.2796 (.1850)	.8201	.7667	.0535

Notes: The probability mean for treatment and control schools are estimated using an ordered logit regression. The treatment effect 'difference' is identified by the marginal change in probability when the dummy variable for treatment changes from 0 to 1. All variables in this table come from student-reported data. Mean probabilities for dummy variables are estimated using a logit regression. All regressions also control for student level variables (e.g. language dummies); teacher and director characteristics (e.g. years of teacher experience) and school and regional level characteristics (e.g. classroom conditions). ***Significant at 1%, **Significant at 5%, *Significant at 10%.