RISE | RICE INITIATIVE *for the* STUDY *of* ECONOMICS

RISE Working Paper 14-019

"Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools"

by

Jere R. Behrman, Susan W. Parker, Petra E. Todd, Kenneth I. Wolpin

RICE
Department of Economics
Baker Hall, MS22
6100 Main Street, Houston, Texas 77005
https://economics.rice.edu

Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools

Jere R. Behrman* Susan W. Parker† Petra E. Todd‡

Kenneth I. Wolpin§

October, 2012

ABSTRACT

This paper evaluates the impact of three different performance incentive schemes using data from a social experiment that randomized 88 Mexican high schools with over 40,000 students into three treatment groups and a control group. Treatment one provides individual incentives for performance on curriculum-based mathematics tests to students only, treatment two to teachers only and treatment three gives both individual and group incentives to students, teachers and school administrators. Program impact estimates reveal the largest average effects for treatment three, smaller impacts for treatment one and no impact for treatment two.

*University of Pennsylvania

†Centro de Investigacion y Docencia Economicas

‡University of Pennsylvania

§Rice University and University of Pennsylvania.

I. Introduction[1]

This paper evaluates the impact of a large-scale social experiment (the Aligning Learning Incentives, or ALI, program) designed to promote mathematics achievement through performance-based monetary incentives. Eighty-eight Mexican Federal high schools with over 40,000 students were randomly allocated to three treatment groups and a control group. Treatment one (T1) provides individual incentives to students only and treatment two (T2) to teachers only. Treatment three (T3) offers both individual and group incentives to students, teachers and school administrators, thus rewarding cooperation among all of the educational actors in the school. The ALI program began in the 2008-2009 academic year and ran for three years, ending in the 2010-2011 academic year. Incentives were determined by student performance on curriculum-based mathematics tests administered to students in grades 10-12 at the end of each year.

An important rationale for utilizing monetary incentives rather than intervening in the educational process directly is that policy makers are not likely to know the best means of improving education given the heterogeneous attributes of students, teachers and administrators across schools. The production process by which school, student and family inputs are transformed into educational outcomes is not well understood.[2] Providing monetary incentives tied to student performance allows students, teachers and principals to choose the best means to improve performance given their circumstances. The ALI program was

[1]The authors worked on this study with the support of the Inter-American Development Bank and the support and collaboration of the Mexican Ministry of Education. We thank these organizations for making this study possible. We are especially indebted to Miguel Szekely, Deputy Minister of Education when ALI was initiated, to Marcelo Cabrol, Santiago Levy and Ana Santiago at the Inter-American Development Bank, to Rafael De Hoyos, Martha Hernandez, Lucia Juarez, Elizabeth Monroy, Araceli Ortega and Diana Zamora who were responsible for the day-to-day operations and management of the project, and to George Wesolowsky who performed the analysis of detecting non-independent test taking.

[2]See for example Hanushek (1986, 2003), Hedges et. al. (1994) and Krueger (2003) for differing views about the interpretation of findings from the education production function literature. Todd and Wolpin (2003) discuss the methodological underpinnings of that literature.

designed to evaluate the efficacy of alternative performance-based incentive schemes.

Research on the impact and efficacy of performance-based monetary incentives in education is relatively sparse. We review some examples from the literature of studies of teacher and student incentives in the next section, restricting attention to randomized control trials in which the measured outcomes are test scores, as in the ALI program. As far as we know, the ALI program is the first randomized control trial to incorporate incentive payments to both students and teachers.[3] Previous studies implement performance incentives for students or teachers only.

A comparison of the ALI program impact estimates to those of prior studies reveals that the treatment effects that are associated with the ALI treatments in which students receive incentives are quite large, especially for the treatment where both students and teachers receive incentives. However, close examination of the test book answer patterns shows that part of the reason for higher test scores in the treatment group is a higher rate of cheating, in the form of student copying, than in the control group, particularly in higher grades and in later years of the program.[4] Our impact analysis therefore provides two sets of estimates of treatment effects, one which does not account for copying and one that adjusts the test scores of students identified to have been likely to have copied a part of their test answers, as determined through a comparison of multiple choice answers of pairs of students and a statistical model for the probability of having matching answers. Even with a liberal criterion for identifying copiers and with two different ways of adjusting their scores, we find substantial program effects on student test scores, indicating that the performance incentive program significantly affected mathematics achievement.

To highlight the magnitudes of the treatment effects, consider the copying-adjusted treatment effect estimates for the 2008-2009 entering 10th grade class. The average treatment effects for the first year they were in the program, as 10th graders, were .17 of a standard deviation for treatment group T1, .01 of a standard deviation for T2 and .31 for T3. In the

[3] There was a recent non-experimental program in Dallas, Texas which paid both students and teachers for passing grades on Advanced Placement (AP) exams (Jackson (2010)).

[4] We do not find evidence of teacher cheating.

second year of the program, as 11th graders, the treatment effects were .30 for T1, .02 for T2 and .44 for T3. Finally, when this cohort reached the 12th grade, treatment effects were .23, .04 and .57. Treatment effects were statistically significant at conventional levels in all three years for T1 and T3, but in none of the years for T2. The pattern of positive effects for T1, even larger effects for T3 and no effect for T2 (with one exception) was also found for all the other 10th grade entry cohorts during the years they were in the program.

A potentially important caveat to our findings is that we cannot ensure that the control students (and to a lesser extent the T2 students), for whom the ALI test is low or no stakes, applied the same level of test-taking effort as the T1 and T3 students. We develop two sets of lower bounds based on alternative assumptions that take into account the possibility of differential effort between the treatment and the control students. In the more conservative and in our view less plausible case, in which the entire differential between T1 and control students is attributed to differential test-taking effort in both years, the T3 effects (adjusted for copying) in year 3 were .31, .17 and .34 for the three grades. In the second case, where only the year one differential between T1 and control students is attributed to test-taking effort, the year 3 effects are, in the three grades, .15, .12 and .13 for T1 and .47, .29 and .47 for T3.

James Heckman has a number of papers illustrating both the value and the limitations of randomized social experiments for policy evaluation purposes.[5] As he shows, randomization provides information only on mean treatment effects for a particular program design without additional assumptions. A central theme of James Heckman's research is that economic models can be used to understand the mechanisms through which a program operates and to investigate effects of program designs that differ in some ways from the program that was implemented. To this end, in section IV, we develop an economic model that can rationalize the results from the ALI experiment.

In the next section, we discuss relevant literature. In section III, we provide details of the ALI experiment, including the overall design, the selection of the sample, a description of

[5] See, e.g., Heckman (2000), Heckman and Smith (1995), Heckman and Urzua (2010)

the tests and the incentive schedules. In section IV, we present the results of the experiment and briefly outline a model that can rationalize the results. Section V concludes.

II. Related Literature

As noted, previous experimental studies of performance incentives provide incentives only to teachers or students.

Teacher Incentives: Muralidharan and Sundararaman (2011) evaluate the effects of teacher performance incentives and school input interventions in rural India. They randomly allocated schools to four treatment groups and to a control group, with 100 schools in each group. One of the treatments was a performance incentive paid to the teacher based on the average improvement in their students' test scores. Tests were administered in mathematics and language at the beginning and end of the school year. The impact of the performance incentive program was a .28 (.17) standard deviation increase in the average test score in mathematics (language).

Glewwe et. al. (2010) conducted a randomized trial over a two-year period that provided primary school (grades 4-8) teachers in 50 rural Kenyan schools incentives based on student performance on district-level exams in 7 subjects. Students in treatment schools had higher test scores (averaged over all subjects) in the second year of about .14 of a standard deviation (.07 s.e.), but the gains dissipated in the third year after the program ended.

Springer et. al. (2010) report on a recent three-year teacher-incentive experiment in the Nashville, Tennessee public schools. Middle school mathematics teachers, who volunteered for the program, were randomly assigned to treatment and control groups. Teachers in the treatment group could earn bonuses, depending on the standardized test scores achieved by their students, of $5,000, $10,000 or $15,000. The magnitudes of the treatment effect estimates in mathematics in grades 6-8 were .06 standard deviations or less and not statistically significant at conventional levels. In grade 5, the average treatment effect was .06 of a standard deviation (.04 s.e.) in year one, .18 standard deviation (.06 s.e.) in year two and .20 (.08 s.e.) in year three.

Student Incentives: Angrist and Lavy (2009) study the effects of a student cash incentive program in Israel that offered high school students incentives for progressing from 10th to 11th grade, 11th to 12th grade and for passing the Bagrut exam.[6] Forty high schools were randomized in or out of the program. The school-based program increased Bagrut passing rates by 6-8 percentage points.

Kremer et. al. (2009) evaluate the impact of a Merit scholarship program for 6th grade girls in Kenya. Schools were randomized to treatment and control groups and scholarships were awarded to the top 15 percent of 6th grade girls in treatment schools based on standardized achievement test scores. On average, girls in participating schools raised their achievement by 0.2 standard deviations. They also estimate impacts by quartile of achievement and find the largest impacts in the second quartile, a group who would seem to have relatively low probabilities of winning an award. The conclusion that learning improved even for those unlikely to win an award is tempered, however, by the finding that the lowest impact is for the lowest quartile.

Fryer (2010) reports results from four different field experiments that implemented various student incentive programs, in Chicago, Dallas, New York City and Washington, DC (in predominantly low-performing urban public schools), with substantial heterogeneity in terms of grade levels participating and incentive design. In New York City, payments were given to fourth and seventh grade students conditional on their performance on ten standardized tests. The program did not yield positive impacts on final year test scores. In Chicago, incentives were paid to ninth graders every five weeks for grades in five courses. The program led to higher grades but had no detectable effect on test scores. The Dallas program gave second graders $2 per book read, with an additional requirement of passing a short quiz on the book, which led to significantly better reading test scores and also to higher grades. Lastly, in D.C., incentives were given to sixth, seventh and eighth graders on a composite index intended to capture their school attendance, behavior and measures of inputs in educational production. The impact estimates are suggestive of substantial positive impacts but are not

[6] The Bagrut exam is a prerequisite for admission to university and for certain types of jobs.

statistically significant. In all cases, a key outcome is student performance on standardized tests. Fryer (2010) concludes from these results that children in these schools do not know what behaviors would lead to improved test-score performance and are thus better served by incentives tied to inputs rather than outcomes.

Levitt et. al. (2010) implement a field experiment to evaluate the effects of a program that gave monthly financial incentives to ninth grade students or their parents in two high schools in a suburb of Chicago. The incentives were given for meeting an achievement standard that is a composite of multiple measures of performance, including school attendance, behavior, grades and test scores. The experimental design varied the award recipient (students or parents) and whether the incentive was awarded piece rate or as a lottery. The incentive amounts had an expected value of \$50 per month, for a total of \$400 per year. The four treatments were randomized at the student level. The study finds modest overall effects of the incentives on achieving the composite achievement standard – a 15-22% increase – but that the effect is not statistically significantly different across treatment types. The study does not find effects of the treatments on standardized test scores.

Barrow and Rouse (2013) evaluate the effect of two performance based scholarship programs for post-secondary students on a variety of outcomes, including student time use, enrollment and GPA. Incentives varied in length and magnitude. The paper finds that students eligible for incentive payments devoted more time to educational activities and less time to work and leisure activities and that the increased effort in their studies was sustained after incentives expired. One puzzling finding was that the students' responses did not appear to depend on the magnitude of the incentive.

It is difficult to generalize from these studies (or from a larger set that includes additional non-experimental studies). First, there are not many evaluation studies of performance-based incentive programs. Second, the existing studies differ in their designs and in the populations studied. And, third, unlike the ALI program, the agent receiving the incentive is not varied within the same population.[7] However, the literature generally finds small measured effects of

[7] Levitt et.al (2010) is an exception, as noted, in which there is randomized variation in whether the

performance-based incentive programs, regardless of whether the student or teacher receives the incentive; an average treatment effect above .3 standard deviations in test scores appears to be unusual.

III. The ALI Experiment

The ALI experiment began with the 2008-09 academic year and ended with the 2010-11 academic year. There were a total of 88 high schools in the experiment, consisting of three groups of 20 treatment schools, each subject to a different incentive design, and a fourth group of 28 control schools with no incentives. In the first program year, there were approximately 12,000 students per treatment and 16,800 students in the control schools. All students in each high school, grades 10-12, participated in the program in each of the three years.

Incentive payments were based on standardized curriculum-based mathematics examinations in grades 10, 11 and 12 given at the end of each academic year. Specifically, the four groups were:

1) Treatment group 1 (T1): Payments to students based on their own performance.

2) Treatment group 2 (T2): Payments to mathematics teachers based on the performance of the students in their classes.

3) Treatment group 3 (T3): Payments to students based on their own performance and on the performance of the other students in their class. Payments to mathematics teachers based on the performance of the students in their classes and on the performance of the students in all other mathematics classes. Payments to non-mathematics teachers and school administrators based on the performance of all of the students in the school.

4) Control group (C): No payments.

student or their parents received the incentive payment.

A. Sample Selection and Characteristics

There are 706 Federal upper secondary schools in Mexico.[8] From that set, we identified 357 schools that were not in their first year of operation and had only one session per day, in the morning.[9] There are four types of Federal upper secondary schools in Mexico: academically-oriented schools, technically-oriented schools with a marine focus, technically-oriented schools with an agricultural focus and technically-oriented schools with an industrial focus. The technical orientation of the latter three school types has diminished over time and all of the schools are now considered college preparatory. Of the 357 "morning-only" schools, 14 fall into the first category, 26 into the second, 183 into the third and 134 into the fourth. Academically-oriented schools have a different mathematics curriculum and so were not included in the ALI program. The schools with a marine specialization also were not included because, after further selection criteria described below, there were too few to randomize across the treatment and control groups.

In addition to federally-administered upper secondary schools, there are also state-administered schools, that are publicly-funded, and private schools.[10] Students successfully completing lower secondary school (9th grade) may apply to any of these schools. Admissions to public (federal and state) and private schools are determined on a competitive basis. To minimize the impact of the ALI program on students' application decisions, and thus on the compo-

[8] Upper secondary schools (high schools) in Mexico encompass grades 10 through 12 and lower secondary schools grades 7 through 9.

[9] The selection of schools was based on data supplied by the Ministry of Education. There were seven schools in their first year of operation, four schools for which the number of sessions could not be determined and 262 schools with multiple sessions. We dropped multiple session schools because each session essentially comprises an autonomous school, having a separate principal (accounting for this, the number of schools is thus about 1000). Clearly, it would have been problematic to have only one session in a given multisession school as part of the program or having different sessions with different treatments. Moreover, given the likely similarities in the student bodies and some overlap in teachers, having both sessions within the same treatment group would reduce the effective number of schools.

[10] In 2008, about 25 percent of upper secondary school students attended federal public schools, 42 percent attended state public schools and 33 percent attended private and other schools.

sition of entering students in years 2 and 3 of the program across the treatment and control groups, schools that were located within 10 miles of another Federal upper secondary school were eliminated.[11] Very small schools (fewer than 200 students) and very large schools (over 2000) were also eliminated, as were schools that had satellite divisions. Finally, schools located in the states of Oaxaca and Michoacan were eliminated due to feasibility constraints. With these restrictions, 135 schools remained out of which 88 were chosen for the experiment. The number of schools was restricted by the need to keep the program on a relatively small scale.

Randomization was performed using a school-based block randomization design. Schools were first grouped into nine blocks, where the block definitions were based on school size and the previous year's graduation rate.[12] Within each block, schools were allocated at random to treatment regimes.[13] The block definitions (cutoffs on school size and graduation rates) were chosen to have roughly similar numbers of schools within each block.

Tables 1 and 2 present evidence on the quality of the randomization among the 88 schools. Table 1 compares the treatment and control schools and the Federal schools not in the experiment on aggregate school-level data supplied by the Ministry of Education. The first two variables, the student population and the graduation rate, were, as noted, used for

[11] Most of the schools are located in rural areas because of the distance criterion. We do not know if there are state schools, autonomous schools or private schools located closer than 10 miles to the ALI schools.

[12] Blocking is a widely-used method for improving the precision of estimated treatment effects by increasing the comparability of the variables used to define the blocks across treatment/control groups. As described in Cox and Reid (2000), the rationale for blocking is to improve precision by using prior knowledge on which baseline characteristics of the units being randomized are likely to be associated with the treatment responses. For maximal efficiency, units should be grouped into blocks so that all units within a block might be expected to give similar responses in the absence of treatment differences.

[13] Following the recommended procedure of Cox and Reid (2000), we randomized six times and chose the randomization where the groups (T1, T2, T3 and C) are most comparable in terms of baseline observed characteristics. Cox and Reid (2000) discuss that if there is any imbalance in observed characteristics that may be related to treatment response, it is better to re-randomize to achieve a better balance in the covariates than to do ex-post regression-adjustment to adjust for imbalance, which would entail a loss in degrees of freedom.

blocking. The other variables in the table were used as additional evidence on the quality of the randomization. They included the following baseline characteristics: percentage of Oportunidades recipients within the school, mean class size, percentages of teachers with university degrees, percentages of new principals, mean distance to the nearest Federal upper secondary school, school type distribution (DGETI or DGETA), and regional distribution.[14] All of the variables in table 1 do not differ from the control group at conventional levels of statistical significance, an indication that the randomization procedure was successful. Table 2 compares treatment and control schools based on the mean score of students on the 9th grade Mathematics ENLACE.[15] This variable was unavailable at the time of the randomization. As seen in table 2, the mean score on the 9th grade ENLACE does not differ between treatment and controls at conventional levels of significance and the largest difference from the mean score of the control group is less than 7 (9) standardized points (.07 (.09) of a standard deviation) for the 10th (11th) grade class.

B. ALI Mathematics Tests

Incentive payments (see below) for all treatment regimes depended on performance on standardized mathematics tests administered at the end of the school year. The tests were designed by CENEVAL based on the input of Mexican experts on upper secondary school mathematics.[16] The tests were based on the curriculum in each grade.[17] The tests were administered by the Ministry of Education with procedures designed to minimize possible

[14] DGETI (Dirección General de Educación Tecnológica Industrial) schools have an industrial focus and DGETA (Dirección General de Educación Tecnológica Agropecuaria) schools have an agricultural focus. Oportunidades is a conditional cash transfer program that provides payments for school attendance to low-income families.

[15] The ENLACE is a national test with separate mathematics and language components that began in 2007 with only two grades and is now administered each year to students in all grades between grade 3 and grade 9 and grade 12.

[16] CENEVAL is a non-governmental organization similar to the Educational Testing Service in the U.S.

[17] The standardized curriculum for each grade is:

Grade 10: Algebra, Geometry and Trigonometry (class hours - 4hrs/wk)

Grade 11: Analytical Geometry, Differential Calculus (class hours -4hrs/wk) testing abuses. In particular, the tests were not administered or monitored by school personnel, but by representatives of the Ministry of Education state offices, with one monitor assigned to each class and an overall supervisor assigned to the school. The same administrators collected the answer sheets and were required to account for all copies of the tests after test administration to reduce the possibility of teaching to the test based on past tests.[18]

For the purpose of determining incentive payments, performance on each test was categorized, as in the 9th grade ENLACE, into four levels: Pre-Basic, Basic, Proficient, and Advanced. A popular method for determining cutoffs for the categories is the bookmark method, which is used for the ENLACE.[19] Using that method with the ALI test scores, in the first program year the percentage of students scoring in the top two categories was zero for the 10th grade, 4.0 percent for the 11th grade and zero for the 12th grade for the treatment and control groups combined. The corresponding percentages for the Pre-Basic category were 76.5, 92.3 and 92.8. This performance reflects the fact that the test design faithfully adhered to the curriculum content, which is quite advanced, especially in light of the low level of pre-high school mathematics skills. Using the bookmark cutoffs would have resulted in few students or teachers receiving incentive payments and almost none receiving the larger payments associated with performance in the top two categories. This result would likely have had a deleterious impact on student and teacher effort in subsequent years of the program. For this reason, the bookmark procedure was not used to establish cut-scores for determining incentive payments. Instead, the cut-scores for the ALI tests were chosen to mimic the 9th grade ENLACE distribution of the control schools for the 10th and 11th grades and to mimic the 12th grade ENLACE distribution of the control schools for the 12th

Grade 12: Probability and Statistics, Applied Statistics (class hours - 5hrs/wk) In 2010-1011, Applied Statistics was replaced with Integral Calculus.

[18] Barlevy and Neal (2011).develop a model of teacher effort choice under an incentive scheme based on the ordinal ranking of students. They show that such a scheme would not require the equating of assessments over time and would thus eliminate the incentive for teachers to "teach to the test." Their framework does not incorporate student effort as a joint determinant of achievement.

[19] See Cizek et. al. (2004).

grade.[20]

C. Structure of Incentive Payments

1. Treatment 1 (Student incentives only)

Table 3 shows the incentive payment schedule for students at each grade level that serves as the basis for Treatments 1 and 3. The amount in each cell represents the payment in pesos for a student with a given level of performance at the start of the grade (the baseline test score as defined above) and at the end of the grade.[21] Payment levels were intended to be large enough to be expected to induce behavioral changes. The payments are similar in magnitude to the attendance incentives given by the Oportunidades program and to a scholarship program offered by the Ministry of Education.

As seen in the table, in the 10th and 11th grades, there was no payment for performance on the ALI tests at the Pre-Basic level. In those two grades, students who scored at the Pre-Basic level on the baseline test received a payment of $4000 pesos if they improved to the Basic level, $9000 pesos if they improved to the Proficient level and $15000 pesos if they improved to the Advanced level. The increments become progressively larger ($4000 for the first increment, $5000 for the second and $6000 for the third), recognizing that the effort necessary to improve from Pre-Basic to Proficient, for example, is likely to be greater than twice the effort in going from Pre-Basic to Basic.

In the 10th grade, students who originally scored at the Basic level received a payment of $2500 pesos for maintaining their achievement level, a smaller amount than those at the Pre-Basic level who improved to the Basic level. The smaller payment reflects the presumably greater effort associated with improvement than with maintenance, a premise that is reflected throughout the incentive schedule. Considerably larger payments were given for improvement beyond the original Basic level, $7500 pesos for achieving the Proficient level and $13500 pesos for achieving the Advanced level. As before, the increments were increasing ($5000 for

[20] The 12th grade mathematics ENLACE was administered for the first time in the 2007-08 academic year.

[21] A dollar was equivalent to about 11 pesos at the time.

the first and $6000 for the second).

Students in the 10th grade who began at the Proficient level received no payment if they fell back to the Basic or Pre-Basic level. If they remained at the Proficient level, they received $6000 pesos, less than that received by students who improved to that level, while if they improved to the Advanced level, they received $12000 pesos. Students who originally scored at the Advanced level received $4500 pesos if they fell back to the Proficient level and $10500 pesos if they remained at the Advanced level.

Bonus amounts were the same for students in the 11th grade with the exception that there was no payment for remaining at the Basic level. The lack of any reward to those in the 11th grade who remained at the Basic level reflected the ALI program emphasis on making progress towards achieving proficiency by the end of the 12th grade. This formulation does, however, lead to an incentive for 10th graders who would perform at the Basic level on the 10th and 11th grade tests to score at the Pre-Basic level on the 10th grade test instead. In that case, the student would receive $4000 pesos in total ($0 in the 10th grade and $4000 in the 11th grade) as opposed to $2500 pesos in total ($2500 in the 10th grade and $0 in the 11th grade). Of course, a student would be uncertain of their 11th grade score and indeed would have been better off scoring at the Basic level in the 10th grade if their score was at the Proficient or Advanced level in the 11th grade. Although the potential incentive incompatibility could have been avoided by giving a bonus of $1500 pesos for remaining at the basic level in the 11th grade, our view was that, given the newness of the ALI test and thus the inherent uncertainties students would have about how well they would perform on the tests (as well as their uncertainty about the cut-offs), the saliency of a zero bonus in emphasizing the goal of attaining proficiency outweighed the potential incentive problem.[22]

The 12th grade payment schedule provided a bonus only for achieving the Proficiency or

[22] A more serious issue would arise, particularly if the program were universally adopted, with respect to performance on the 9th grade ENLACE, where it would be unequivocally better for a student faced with the prospect of 10th grade ALI incentives to have peformed at the Pre-Basic level However, this incentive would be mitigated by the fact that the 9th grade ENLACE is a high stakes test used by high schools as part of their competitive admissions criteria.

Advanced levels of performance, reflecting the goal that students reach at least the proficient level by the time they graduate. There was a significantly higher payment for performance at the Advanced level, \$10000 pesos, as opposed to the Proficient level, \$5000 pesos.

2. Treatment 2 (Teacher incentives only)

In treatment 2, mathematics teachers were rewarded for the performance of the particular students they taught during the year. The reward was based on the sum of the rewards earned by the students as described in table 4. The per-student bonus was 5 percent of the bonus payments in the student schedules, except for the modification that teachers were penalized for students in the 10th and 11th grades who were not at the Proficient or Advanced levels at the end-of-grade test and who performed more poorly on the end-of-grade test than on the baseline test.

Consider, as an example, a teacher who had a 10th grade mathematics class. For each student that improved from the Pre-Basic to the Basic level, the teacher would receive 200 pesos (5 percent of the \$4000 pesos that such a student would earn for him/herself). If such a student instead improved to the Proficient category, the teacher would earn \$450 pesos (again 5 percent of the \$9000 pesos the student would receive). If, however, a student lost ground, for example, moving from the Basic to the Pre-Basic level, the total of the student payments used to calculate the teacher reward would be reduced by \$125 pesos. A teacher's total payment was bounded below by zero.

A teacher with an 11th grade mathematics class faced the same incentive schedule as for a 10th grade class except that, in conformity with the student incentive schedule, there was no payment for a student whose starting and ending test score was at the Basic level. A teacher with a 12th grade class received \$250 pesos for each student who reached the Proficient level and \$500 pesos for each student that reached the Advanced level (5 percent of the student payment).

Schools operate on a semester basis. To obtain the academic year (two-semester) sum, each student that the teacher had in their class during a semester was counted as one-half a student. The total earned by a teacher was the sum of the earnings from all of the students in that teacher's classes over both semesters. To get an idea of the magnitude of the payments, consider a full-time teacher, one teaching five classes in each semester, who had 200 (year-equivalent) students (an average class size each semester of 40 students). Suppose that the average student payment was $2,500 pesos and that no student fell back. Such a teacher would earn $25,000 pesos, which is a bonus of between 10-15 percent of the annual salary of a teacher in a Federal high school.

A specific aim of the teacher incentives design was to avoid teachers concentrating their effort on high-performing students.[23] It did so in three ways. First, teachers gained more from a lower-performing student achieving a given level than from a higher-performing student achieving that same level; for example, a teacher with a 10th grade class earned $200 pesos if a student who scored initially at the Pre-Basic level improved to a Basic level, but only $125 pesos for a student who scored initially at the Basic level and remained there. As with the students, the teacher effort required to elicit an improvement in a student initially scoring at the Pre-Basic level is presumably greater than the effort to maintain the score of the Basic-level student. As seen in table 4, the $75 peso differential between initial Pre-Basic and Basic test scores carried over to the other final test categories. It also carried over between any two adjacent initial test score categories, that is, between Basic and Proficient and between Proficient and Advanced. This pattern resulted in a doubling of the differential between a student initially scoring at the Pre-Basic level and improving to the Proficient level and a student initially scoring at the Advanced level and falling back to the Proficient level ($450 pesos vs. $225 pesos).

Second, and opposite to the "carrot" that compensated teachers for the extra effort associated with improving scores of low-performing students, the $125 peso penalty incurred if a student regressed acts as a "stick" aimed at maintaining students at least at their

[23]Neal and Schanzenbach (2010) use data from No Child Left Behind (NCLB) to analyze how teachers may have incentives to concentrate on subsets of students, in their case, students near cut-off values that determine whether school-level goals are met.

initial performance level. Third, for students initially at the Pre-Basic level, the possible payments to the teachers (and to the students) were strictly non-negative, with relatively large payments if the students improved a great deal. For example, the mathematics teacher received $200 pesos if a student who started at Pre-Basic in the 10th grade advanced to Basic, $450 pesos if a student who started at Pre-Basic advanced to Proficient, and $750 if a student who started at Pre-Basic advanced to Advanced.

In contrast to the ALI design, in an incentive system that depended on, say, the average performance of teachers' students, teachers would receive a greater reward if low-performing students were encouraged to not take the examination. Under the ALI design, that was not the case for the lowest performing students; those who scored Pre-Basic in the baseline test could not subtract from the teacher's reward and, as noted, could contribute considerably if their performance improved in the end-of-year test. For students with Basic and Proficient baseline scores, there was some potential for teachers losing if the students dropped back due to the "stick", but the "carrot" was intended to be sufficient to offset that potential loss.

3. Treatment 3 (Aligned student, teacher and administrator incentives)

Students: In Treatment 3, in the 10th and 11th grade, each student received a reward based on individual performance as in Treatment 1 (according to the schedule in table 4) and also on the performance of the other students in his/her mathematics class. In the 12th grade, the student received a reward based only on individual performance. The first component was calculated in exactly the same way as in Treatment 1. The second component was calculated as a fixed proportion, one percent, of the total payments earned by classmates.[24]

The rationale for paying students for the performance of their classmates rests on possible synergies of two different kinds, both of which depend on there being a fundamental complementarity between a student's own effort and classmates' efforts. In one case, the effort

[24] In calculating the class sum over a year's time, the class sum in each of the two semesters is multiplied by one-half and then added together. This procedure accounts for compositional changes in classes, that is, that a student may not be in classes with the same students in both semesters.

of one's classmates is a pure externality in which a positive climate or culture of learning can be created by the overall effort within the classroom. This climate affects the amount of effort each student puts into their own learning. A second synergy arises when students actively help other students, for example, if higher-performing students tutor lower-performing students. The component of the bonus payment based on class performance provided an (extra) incentive for this behavior. Not only may such activities improve learning among low performers, they may also improve learning among high performers as teaching itself can lead to a deeper understanding.

Mathematics Teachers: The reward to a full-time mathematics teachers was the sum of the total performance payments earned by the students in their classes calculated as in Treatment 2 (according to the schedule in table 5) and a fixed proportion, 25 percent, of the average full-time equivalent adjusted performance payments earned by the other mathematics teachers (across all grade levels).[25] The rationale for the second component of the reward was to stimulate cooperation among the mathematics teachers. Such cooperation may take the form of formal or informal discussions of teaching methods and subject matter, mentoring less experienced teachers and directly sharing lesson plans or other class materials.

Non-mathematics Teachers: Non-mathematics teachers received a cash payment that is 25 percent of the school-wide average (full-time equivalent) mathematics teacher performance payment. Payments for part-time teachers were adjusted for their own full-time equivalence status. The rationale for this payment recognizes the potential importance of the overall learning environment in the school and of the potential value of interactions among teachers from different disciplines in sharing ideas about pedagogy (and perhaps subject matter in the case of allied fields like physics) and about students that they have in common.

Principals and Associate Principals: Principals received a cash payment that is 50 percent of the average full-time equivalent mathematics teacher performance payment. Associate principals received a cash payment that is 25 percent of the school-wide average full-time equivalent mathematics teacher performance payment, adjusted for their own full-time

[25] Rewards to part-time teachers were pro-rated by their own full-time equivalence status.

equivalence status. These payments recognize the importance of support services provided by administrative personnel in fostering learning within the school.[26]

IV. Results

A. ALI Test Completion Rates

Most students in Mexico who complete 9th grade enter high school (10th grade). Of the cohort that entered kindergarten in 1996, about 65 percent completed 9th grade and 95 percent of those continued to 10th grade. However, only 76 percent of those who entered 10th grade actually completed the grade and, of those, 81 percent graduated from high school. Given high attrition rates at the national level, one might expect to see a significant proportion of students in the ALI schools who begin the school year but do not take the ALI examination at the end of the year. To the extent that attrition is not uniform across the treatment and control schools, treatment effect estimates based on simple mean comparisons could be biased.

Table 5 provides attrition figures for the 2008-09 10th grade cohort, both for enrollment between the first two (year 1) semesters and for the completion of the ALI tests over the three years. The first panel of figures show the percentage of students in treatment and control schools enrolled in the Fall semester who were also enrolled in the Spring semester. Attrition between the Fall and Spring semesters is 9.9 percent for the controls and 12.0, 9.5 and 13.2 percent for T1,T2 and T3; none of the differences are statistically significant at conventional levels. The fact that continuation rates do not differ significantly either in magnitude or statistically across the treatment and control groups may be surprising, particularly for T1 and T3 where students receive direct monetary incentives. Indeed, one might have expected the ALI program to have reduced dropout rates. However, as already noted, ALI is not the only program providing attendance incentives. Almost 40 percent of the students in ALI schools receive a substantial attendance subsidy as part of the Oportunidades conditional

[26]The formulae used for calculating the bonuses for students, mathematics teachers, non-mathematics teachers, principals and associate principals are given in Appendix B.

cash transfer program. In addition, as part of another scholarship program, students whose family income is below the poverty line and who successfully progress from one grade to another receive a scholarship payment; the baseline scholarship amount is considerable with increments for achieving a high grade point average. Given these subsidies already in place in all Federal high schools, it is less surprising that there is no discernible additional effect of the ALI program on the dropout rate between semesters. The existence of these additional programs does not pose a problem for the estimation of the marginal effect of the ALI program, although it could affect the generalizability of the ALI program impact results to other settings.[27]

As also seen in table 5, conditional on enrollment in both semesters, 85.9 percent of the controls took the ALI exam, about the same as for T2, but about 3 percentage points lower than for T1 and T3; again, the differences from the control group are not statistically significant. This pattern of test-taking is maintained in years 2 and 3. For example, 67.8 percent of the controls who took the ALI exam in year 1 as 10th graders also took the exam in year 3 as 12th graders. The comparable figures for T1, T2 and T3 are 70.9, 66.4 and 70.3 percent, which do not differ statistically from the controls. We previously noted that there is potentially an incentive for teachers in T2 and T3 schools to try to identify students in the 10th and 11th grades that would do worse on the ALI test than on the baseline 9th grade ENLACE test and in some way have them not take the ALI test. However, T3 students cannot lose anything by taking the test, which acts to counterbalance the teacher incentive. Moreover, the poorest performing students on the ENLACE test, those that scored at the Pre-Basic level, cannot fall back so there is no incentive for the teachers in T2 or T3 to discourage them from taking the test.[28] Given the data in table 5, it does not appear that

[27] Recall from table 2 that the percentage of Oportunidades recipients did not differ between the treatment and control schools. Student self-reports in year 3 indicate that the percentage of students receiving a Minister of Public Education scholarship ranges between 11 and 13 percent across the treatment and control groups.

[28] In addition, teachers are not officially informed of student 9th grade ENLACE scores, which is relevant for determining teacher bonuses for the 10th grade (and for the 11th grade in the first year), nor are they

teachers had manipulated the test-taking sample in any appreciable way.

B. Treatment Effects

As we discussed, the protocol for the administration of the ALI test, which called for an external monitor in each classroom, was intended to minimize the possibility of cheating. To determine whether and to what extent cheating had occurred, an independent statistical analysis of student answer sheets was conducted by George Wesolowsky, based on the methods described in his Wesolowsky (2000).[29] The method specifies a statistical model for the probability that student i correctly answers multiple choice question j, incorporating a parametric function of the "difficulty" of the question and the "ability" of the student. Using that model, it is possible to determine, for every pair of students and for each question, the probability that two students will have the same answer assuming that all wrong answers are equally likely. Choosing a critical value for the number of observed matches, the null hypothesis of no copying was tested for each student pair. The critical value was based on a Bonferroni correction such that the probability was one that at least one pair of students would be falsely accused. Given that criterion, we interpret the amount of cheating that is identified as an upper bound.[30]

In each pair, we assigned the student with the higher 9th grade ENLACE score as the

informed about the ALI test score results of individual students, which is relevant for determining teacher bonuses for the 11th grade.

[29] To our knowledge, student copying has not been studied in relation to student incentive performance programs. However, the education literature demonstrates that a significant fraction of students admit to cheating during regular school examinations (Cisek (1999)).

[30] The analysis compared pairs of students within a grade independently of treatment group, school or class. The number of pairs identified who were not in the same treatment group or school was always negligible, supporting the validity of the method. However, that was not always the case across classes, indicative of the relatively low critical value used to determine cheating pairs. In particular, 6.0 percent of cheating pairs crossed class bounderies in year 2 for grade 12, 9.8 percent crossed class boundaries in year one for grade 12, 14.4 percent crossed class bounderies in year 3 for grade 11 and 29.2 percent for grade 12 in year 3. Unfortunately, we cannot rule out that in some cases classes were combined for the test administration, though no such occurrences were reported by the overall external supervisor assigned to the school.

source of the answers and the other student as a "copier."[31] If a student is ever a copier in any pair, even if the student was a source in another pair, the student's final designation is as a copier. Table 6 provides the results from the cheating analysis. The table reports, for each grade and programmatic year, the percentage of students who were members of a cheating pair and the percentage who were copiers.[32] The estimated (upper bound) percentage of copiers in the control group varied between 2 and 6 percent, depending on grade and year, with no obvious pattern across grades or years. The extent of copying was similar for students in T2, who, like the controls, had no direct ALI incentive. However, the percentage of copiers was considerably higher and of similar magnitudes in both T1 and T3, the treatments for which there were direct student incentives. The percentage of copiers was especially high for the 11th grade in years 2 and 3 and for the 12th grade in year 3; the highest percentage of copiers was for the 11th grade in year 3, 19.8 percent for T1 and 23.8 percent for T3.[33]

As noted, teachers did not administer the test nor handle the test booklets, and all test copies were to be returned to the state ministry offices. The finding that the extent of cheating was no greater in T2 than in C nor greater in T3 than T1 is consistent with that protocol being followed.[34] In addition, we analyzed the difference in scores between treatment and control schools based on the "anchor" questions, that is, the 30 percent of

[31] To determine the accuracy of this classification, we compared the difference between the 9th grade ENLACE and the ALI test scores for three groups: non-cheaters, sources and copiers. In essentially all years, grades and treatment groups, the difference in scores for copiers was far greater than for the other two groups which were themselves similar. For example, in year 2 among the controls the difference between the scores was -34 for non-cheaters, -40 for sources and 53 for the copiers. The similar figures for T3 were 8.9, 21.2 and 190.8. In defining copiers, for cases in which the 9th grade ENLACE was missing, the assignment was based on the ALI test score.

[32] The average percentage of copiers is more than half the percent of all cheaters because some students were sources for more than one copier.

[33] Cheating was highly concentrated among a few schools. In many cases, the top three schools accounted for three times as many copiers in the treatment groups in a given year and grade as the schools accounted for the total number of students. At the extreme, for example, the top 3 T1 schools in year 2 for grade 11 accounted for 73.3 percent of the copiers and only 21.7 percent of all of the students.

[34] For an analysis of teacher cheating see Jacob and Levitt (2003).

questions that were repeated each year.[35] Larger treatment effects based on those questions alone might indicate that teachers somehow gained access to the tests and used them "to teach to the test" in subsequent years. However, we found the treatment effects (either adjusted for cheating or unadjusted) to be no larger for the anchor questions.

Given the results of the cheating analysis, we report average treatment effects in table 7 both with and without an adjustment to the ALI scores of the copiers. The treatment effects are based on standardized test scores, normalized (as in the ENLACE) to have a mean of 500 and a standard deviation of 100 for the controls. The predicted ALI test scores for copiers in grades 10 and 11 are based on regressions of the ALI scores of the control students (who were not copiers) on their 9th grade ENLACE score and a dummy variable for a missing ENLACE score; for the 12th grade, the regression included in addition the student's 12th grade ENLACE score and a missing dummy.

There are five 10th grade entry cohorts that are observed in different grades over the three years of the ALI program. For example, the 2006-07 10th grade entry cohort was observed only once, in the 12th grade in year 1 and the 2007-08 cohort was observed twice, in the 11th grade in year 1 and in the 12th grade in year 2. The samples for each cohort shown in table 7 are based on of non-attriters over the years each cohort is observed.[36] The average treatment effect is the difference in the mean test score of each treatment and the control group conditioning on the following pre-program school-level variables to increase precision: the school-wide average 9th grade mathematics ENLACE scores for the current 10th, 11th and 12th grade students, the 2007-08 academic year (AY) average 12th grade mathematics and language ENLACE scores, regional dummies and state dummies (for the five states in which there was at least one school in each ALI group). Standard errors account for clustering at the school level. The estimates with the adjustment represent, in our view, a reasonable lower bound on treatment effects.[37]

[35] Anchor questions are not supposed to be more or less difficult than other questions.

[36] The results are not sensitive to the exclusion of attriters. As discussed, attrition was similar across the treatment and control groups.

[37] In Appendix C, we report results in which the ALI scores of copiers were set equal to the 25th percentile

In year one, as seen, the adjusted and unadjusted treatment effects are very close for grade 10. Based on the adjusted estimates, T3 students in the 10th grade scored 31.4 standardized points higher, on average, than did C students and T1 students 16.9 standardized points higher.[38] A 95% confidence interval for T3 ranges from 19.8 to 43.0 points and for T1 from 7.1 to 26.7. The score for T2 students is almost identical to that of C students (the point estimate is 1.27 points with a s.e. of 5.74). The p-value for the (two-sided) test that the T3 treatment effect differs from the T1 effect is .010 and that for the test of T3 against T2, less than .001. The adjusted and unadjusted estimates diverge slightly for grade 10 in year 2. Using the more conservative adjusted estimates, the T3 treatment increases test scores by 46.6 standardized points over the controls and the T1 treatment by 29.1 points. As in year 1, the T2 treatment has no effect on test scores. Both the T3 and T1 effects have relatively narrow 95 percent confidence intervals and the difference between the T3 and T1 is statistically significant. A larger difference between the adjusted and unadjusted treatment effects emerges in year 3, with the adjusted effects being about 25 percent lower. The adjusted treatment effects are 63.4 points for T3, 32.3 for T1 and 13.5 for T2 (the only time that there is a statistically significant T2 effect across all grades and years).

There are a number of reasons why the program effect, as seen, would be expected to grow over time. First, the schools were not informed of the ALI program until well into the first semester of the first year and it took some time after that for the students to be apprised of the program details. Second, the only information given to the students and teachers about the test was that it would adhere strictly to the curriculum of each grade. In the first year, there would be considerable uncertainty about the format and difficulty of the test for both students and teachers. Having experience with the test in the first year would resolve some of that uncertainty, even though the tests themselves were not left behind to minimize the possibility of "teaching to the test."

---

score of the control group students who were non-copiers. The results do not differ much for the tenth grade in all years and for the 11th and 12th grades in year 1. The magnitudes of the treatment effects for the upper two grades in years 2 and 3 are somewhat smaller than those reported in table 7.

[38] Each standardized point represents .01 of a standard deviation.

The eleventh and twelfth grade results in many ways mimic those of the tenth grade. The adjusted treatment effects are larger in years 2 and 3 than in year one and are of similar magnitude as those for the tenth grade. Estimated treatment effects are larger for T3 than for T1, each is precisely estimated, and they are statistically distinguishable. There is no discernible T2 effect. In year two (three), based on the adjusted estimates, T3 students in the 11th grade scored 43.7 (42.1) standardized points higher, on average, than did the C students, and 12th graders 34.8 (56.7) points higher.

1. Heterogeneous Treatment Effects

Gender: Table 8 reports both adjusted and unadjusted average treatment effects for the 2008-09 tenth grade cohort for each of the three years of the program by gender. Recall that this sample consists of students who took the ALI test in all three years. There is little gender difference in average treatment effects at any grade.

Ninth Grade ENLACE: Tables 8 also reports average treatment effects for this cohort distinguished by performance on the 9th grade ENLACE. Recall that the incentive schedule was designed so that low-performing students would have a larger incentive to improve their mathematics knowledge, owing to the presumed greater effort necessary to achieve any standard, and teachers would not have an incentive to specialize their efforts on higher performing students. Table 8 provides evidence on whether the incentive design accomplished these goals.

ALI incentives induced significant increases in test scores across all 9th grade ENLACE performance categories: Pre-Basic, Basic and Proficient and Advanced combined. Concentrating on the adjusted estimates, the T3 treatment effect for those students in the Pre-Basic ENLACE category was 26.8 points in the tenth grade, 33.4 points in the 11th grade and 50.7 points in the twelfth grade. The comparable effects for T1 were 15.0, 24.4 and 23.6. Those students scoring in the Basic category on the ENLACE scored higher than students in the Pre-Basic category in the 10th and 11th grades, but about the same in the 12th grade. Students scoring in the two highest ENLACE categories have the highest treatment effects, 28.0, 47.3 and 45.6 for T1 in the three grades and 45.3, 58.1 and 70.2 in T3 for the three grades. T2 effects for all ENLACE categories and in all grades are small, in a few cases negative, and not statistically significant.

2. Potential Caveats

Test-taking effort: The estimated treatment effects for T1 and T3, even those that are adjusted for copying, appear to be large relative to the literature, especially for T3 in year 3. An implicit assumption is that the test-taking effort of the students in the control schools (and to some extent the T2 schools) is the same as that of the students in the T1 and T3 schools. Given that the ALI test is a low (if not a no) stakes test for the C students, it is possible that the average treatment effects are exaggerated. Some evidence can be brought to bear on the issue.

We use two ways to obtain a lower bound on the treatment effects that accounts for differential test-taking effort, in one case a lower bound only for T3 and in the other for both T1 and T3. The first way assumes that the test-taking effort of T1 students is no less than that of T3 students. Although there are greater monetary incentives for T3 students and teachers due to group incentives, and administrators are also rewarded, it seems reasonable to assume that the incentives for the students in T1 are substantial enough for students to want to maximize their effort on the test. Under that assumption and also assuming that the *entire* difference between the test outcome of T1 and C students in each year is due to a lack of test-taking effort on the part of the controls, the difference in the treatment effect between T3 and T1 would be a lower-bound estimate of the treatment effect for T3. The results in table 7 (with the copying adjustment) imply that, for year three, the lower bound for the treatment effect for T3 is 31.1 (63.4 - 32.3) standardized points for grade 10, 16.9 for grade 11 and 33.9 for grade 12.

The second way assumes that test-taking effort of the students in the control schools is at least as great in year three as in year one and, further, that the treatment effect for T1 in year one was *entirely* due to the lack of test-taking effort of the controls, that is, that the true treatment effect in year one was zero for T1. In that case, a lower-bound estimate of the treatment effect is the difference between the treatment effect in year three and the treatment effect of T1 students in year one. This assumption leads to the following lower-bound estimates of treatment effects (adjusted for copying) in year 3 for T1 of 15.4 for the 10th grade, 11.6 for the 11th grade and 13.1 for the 12th grade. Similarly, the lower-bound estimates for T3 are 46.5 for the 10th grade, 28.5 for the 11th grade and 47.0 for the 12th grade.

Clearly, the first bounding approach is more severe in terms of the influence of the lack of test-taking effort on treatment effects, implying as it does a zero treatment effect for T1 in all three years and thus, given the increased performance of T1 students over time, a sizable reduction in test-taking effort of the controls in year three relative to year one. Perhaps one could make an argument that the 12th graders, who, in year three, were taking the ALI test for the third time, and the 11th graders for the second time, had diminished test effort, but such an argument is less compelling for the 10th grade students in year three who were taking the ALI test for the first time. Indeed, raw scores (the percentage of questions answered correctly) for the control school students were similar in all three years in any of the grades, which is consistent with a constant level of test-taking effort as assumed in the second set of lower bound estimates.

12th grade ENLACE: In the middle of the spring semester of the 12th grade, students are administered a national mathematics "competency" test. The scores of the students in the ALI schools, which were made available to us, can be used to see whether the effects of the treatments on the curriculum-based ALI tests carry over to the 12th grade ENLACE. Based on the same regression specification as in table 7, the treatment effect estimates for the 12th grade students in year 3 of the ALI program were (standard errors in parentheses): -19.1 (9.12) for T1, -15.6 (9.12) for T2 and 18.3 (13.7) for T3. These results are not only quite different from those for the ALI exam, but the negative and statistically significant effect for T1 (given the randomization) is anomalous. Whether these results are due to the very different content of the ALI and ENLACE tests or perhaps to diminished test-taking effort, particularly for the T1 and T3 students who are concentrating their effort and attention on the end-of-year ALI tests, is unclear.

Translation to raw percentage scores Measured in standard deviation units, the treatment effects estimated under the ALI experiment are large relative to the range of estimates reported in the experimental incentives literature. However, as measured by raw percentage scores, the performance of the treatment groups is less striking. For example, for the 2008/09 cohort, students in T3 (T1) answer only 45.3 (43.6) percent of the questions correctly on the 10th grade test as compared to 41.7 percent for the controls. Similarly, on the 11th grade test, the average score for the T3 (T1) students is 42.8 (41.3) verses 38.2 for the control students, and, on the 12th grade test, the T3 (T1) score was 48.4 (45.1) as opposed to the control average score of 42.9. As previously noted, the test covered the curriculum for each grade. Although students in T3 and T1 gain relative to controls, in absolute terms they master less than half of the curriculum.

3. Student and Teacher Effort

As part of the ALI project, students and teachers participating in the ALI program were given (self-administered paper) questionnaires that attempted to measure learning and teaching effort in each semester of each year. Table 9 compares the survey answers in year 3 across the control and treatment groups, for the students who took the ALI test and responded to the survey and whose teachers also responded to the survey (which is true for about 90% of the students who took the exam). The estimated treatment-control differences are broadly consistent with students in T1 and T3 groups having higher levels of effort than those in the control group and with students in the T2 group behaving very similarly to the controls. For example, students in T1 and T3 spent more time studying mathematics, were significantly less likely to text or watch TV while doing homework, were significantly more likely to give help to classmates, and reported putting "much effort" into their school work. Students in T1 and T3 also reported spending no less time studying subjects other than mathematics, which implies that the ALI incentives did not appear to have shifted study effort away from other subjects.

Table 9 also compares measures of teacher effort. The evidence is mixed on whether teachers in the treatment groups exerted more effort. A higher fraction of teachers in the treatment groups reported preparing their students for the ALI exams and helping their students outside of class to prepare for the exam, with the highest fractions observed in the T3 group. However, there were a number of measures related to time spent preparing for class (not shown in the table) where no difference was observed between the different groups of teachers.

C. Rationalizing the Results with a Model of Student and Teacher Effort Choice

A perhaps puzzling feature of the results is the large differential between T3 and T1 effects and a zero effect of T2. One potential explanation is that T3 is not simply a combination of T1 and T2 given that T3 includes bonuses for students and mathematics teacher based on the performance of peers and for non-mathematics teachers and administrators. It should be noted, however, that the bonuses based on peer performance account for only about 25 percent of the total bonus for both students and teachers. Given the large differential between T1 and T3 treatment effects in many cases, it may seem implausible that it is mostly due to the additional elements in T3 beyond the incentives for student and teacher own performance.

It is useful to ask whether this pattern of results can be generated by a model of the determination of student performance. To that end, we consider a strategic model of the effort choices of the students in a class and of the teacher. Each student, s, begins a grade with a given level of knowledge, denoted by k_{0s}, and the teacher, t, with a given stock of instructional capital, k_{0t}. During the school year, each student supplies learning effort, ε_s, and the teacher supplies instructional effort, ε_t, which is a public input. End-of-year knowledge, K_s, is given by the production function

$$K_s = F(\varepsilon_s, \varepsilon_t; k_{0s}, k_{0t}, S) \qquad (1)$$

where S is the size of the class. Students care about their end-of-year knowledge and teachers about the knowledge of the students in their class. Students and the teacher face effort cost functions $c_s(\varepsilon_s)$ and $c_t(\varepsilon_t)$. Each student in the class maximizes

$$V_s = U_s(K_s) - c_s(\varepsilon_s), \tag{2}$$

where $U_s(K_s)$ is the utility the student receives from end-of-year knowledge, and the teacher maximizes

$$V_t = U_t(\mathbf{K}_s) - c_t(\varepsilon_t), \tag{3}$$

where $\mathbf{K}_s$ is the class vector of student end-of-year knowledge (determined by (1) for each student).[39] Student and/or teacher bonuses can be accommodated in the model by augmenting the student and teacher utility functions to include the ALI bonus schedules.

Todd and Wolpin (2012) develop sufficient conditions, in which students and the teacher play a Nash game, that can generate the ALI results and an estimation method to empirically implement the model. The critical assumptions are (1) the student can supply a minimal level of effort at zero cost but must pay a fixed (and variable) cost for supplying effort above that minimal level; (2) the marginal product of teacher's effort is zero if the student chooses to supply only minimum effort and (3) student and teacher effort are complementary inputs in (1). Given these assumptions, a teacher bonus alone may not be sufficient to induce enough students within a class, who without the bonus were supplying minimum effort, to supply above-minimum effort in response to an increase in teacher effort (given complementarity of student and teacher effort). A student bonus alone, given it directly affects student incentives, can induce such a response and will also increase teacher effort (given complementarity). Once a student bonus is in place and students are supplying above minimum effort, an additional teacher bonus can further augment both teacher and student effort. It is thus possible to observe that the T2 - C difference is zero, and both the T1 - C effect and the T3 - T1 effect are positive. Alternatively, it is possible to generate the ALI results through restrictions on the shape of the production function. Such an explanation would seem to require a significant degree of non-convexity (akin to a fixed cost).

[39] One also could allow the teacher and/or student utility to depend on initial knowledge or value-added.

Within any class, the solution of the model depends on the teacher's preference over each student's end-of-year knowledge, the teacher's instructional capital and effort cost and on the distribution within the class of student initial knowledge, preferences over knowledge and effort costs. In the ALI project, as noted, we collected information from surveys of teachers and students that provide measures and determinants of these characteristics across all of the schools and classes. Todd and Wolpin (2012) exploit that data to estimate a version of the model that allows for heterogeneity across schools, teachers and students in these attributes. Extrapolating the ALI results to other populations in which the distributions of these attributes differ would require specifying how they differ and related data to account for those differences. The model provides the framework for the data collection that would be necessary.

D. The Cost of the Treatments

Table 9 summarizes the program cost in terms of the incentive payments for the second year and for each of the three treatments.

Treatment 3: Tenth grade students in T3 on average earned \$2,991 pesos based on their own performance and an additional \$1,108 pesos based on the performance of the students in their class. Thus, the class contribution was about 25 percent of the total. Eleventh grade students earned \$2,679 pesos on average from their own performance, lower than what 10th grade students received primarily because there was no payment for students who performed at the Basic level on the baseline ENLACE test and on the ALI test. They earned an additional \$861 pesos from the performance of their classmates. Twelfth graders earned on average \$991 pesos, less than the other grades because their payment was contingent on scoring at least at the Proficient level; there was no additional payment for the performance of classmates.

A full-time mathematics teacher received on average \$15,330 pesos for the performance of the students in their classes and an additional \$3,779 pesos for the performance of the students taught by the other mathematics teachers in the school. Each full-time equivalent non-mathematics teacher or assistant principal received \$3,872 pesos and the principal received \$7,744 pesos.

Taken together, the cost per student was \$3,303 pesos (about 275 U.S. dollars), about 15% of the current per student expenditure in Federal upper secondary schools.

Treatment 1: Tenth grade students in T1 received \$2,515 pesos, eleventh grade students \$2,541 pesos and twelfth grade students \$915. The cost per student was \$2,080 pesos, about 10 percent of the current per student expenditure.

Treatment 2: A full-time mathematics teacher received on average \$6,332 pesos for the performance of the students in their classes. The cost per student was \$43 pesos.

A part of the incentive payments of the ALI program is "wasted" in the sense that some students were paid for results that would have been achieved without the incentive. It is possible to compute the magnitude of the waste based on the payments the C students would have received under the different treatments using the transition rates between the ALI score categories for the C students. For example, of those C students who scored at the Pre-Basic level on the 9th grade ENLACE, 24.0 percent jumped at least one category on the 10th grade ALI test, thus earning a reward in the hypothetical case that they had been in the T1 or T3 treatment. Of course, had they been in one of those treatments, our estimates indicate that many more would have jumped to a higher category and some of those that jumped one category would have instead jumped two or three. Nevertheless, these students would have been rewarded, in at least some part, for an outcome that would have been achieved without the treatment. As seen in table 9, the waste is substantial, amounting to 49.7 percent of the cost of T3 and 55.9 percent of the cost of T1. Although one would prefer a program in which the waste was small, what matters for assessing the efficacy of the program is a comparison of the program's benefits relative to its costs.

V. Conclusions

This paper evaluates the effect of the ALI pilot program that randomly assigned 88 Mexican high schools to three treatment groups and to a control group to measure the effectiveness of three alternative performance incentive schemes on mathematics tests scores: (T1) incentives for students only, (T2) incentives for teachers only, (T3) individual and group incentives for students, teachers and administrators. Previous studies used randomized trials to analyze effects of student only or teacher only incentive schemes, but the ALI program is the first experimental study to combine student and teacher incentives.

An analysis of student test scores finds very large treatment impacts for treatments (T1) and (T3) that pay incentives paid to students. Further examination of the test score answer booklets revealed that the student incentive payments also induced a higher rate of cheating in the form of student copying. We proposed an adjustment procedure to account for the higher rates of copying estimated to have occurred in these treatment groups vs. the control group. Even after making adjustments for copying, we find substantial program impact estimates on mathematics achievement.

Our general findings, based on the copying adjustment, can be summarized as followings: (i) Providing the ALI incentives to students alone increased mathematics test scores by 0.2-0.3 of a standard deviation, depending on the grade and year (ii) Providing ALI incentives to teachers alone did not affect test scores, (iii) Providing ALI incentives to students and mathematics teachers (both for their own performance and for that of their peers', and for other teachers and school administrators) led to the largest treatment effect estimates, increasing test scores by 0.3-0.6 of a standard deviation and (iv) Analysis of treatment effects conditional on initial test score performance categories shows that there are positive impacts across the entire baseline test score distribution. Lastly, our sensitivity analysis explored the robustness of the impact estimates to allowing for differential test-taking effort between the treatment and control groups. Assuming, for example, that all of the measured T1 impact in the first year is due to test taking effort, we still find large impacts of treatment T3.

The fact that incentive payments in the ALI program are performance-based means that the cost of any treatment increases with its success. The per student cost of T3, the most successful treatment, is 50 percent more than T1. However, T3 and T1 differed in many ways, with T1 providing rewards based on individual student performance only and T3 rewarding both individual and group performance for students, teachers, and administrators. We can isolate the cost of each of the components of T3, but the experimental design does not identify the relative contribution of each component to T3's greater success. Similarly, the cost of T2 was less than 1.5 percent of the cost of T3, but T2 was found to be ineffective.

A limitation of the ALI experiment, as of all experiments, is that we can only learn the impact of the treatments that were tried and only in the population studied. Extrapolations to other populations would have to account for the fact that the subset of Mexican federal high schools studies was selective and that there were preexisting programs that provided school attendance subsidies. To learn about effects of other treatments, such as variations in the performance incentive schedules, requires either additional experiments or the development and estimation of behavioral models that can be used to extrapolate to other hypothetical treatments or to other populations.[40]

[40] See Todd and Wolpin (2006) for an example of the use of experiments and modeling for the purpose of performing counterfactual experiments.

References

[1] Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *The American Economic Review* 99: 1384-1414 (31).

[2] Andrews, Donald W. K., and Moshe Buchinsky. 2000. "A Three-Step Method for Choosing the Number of Bootstrap Repetitions." *Econometrica*. 68:23-51.

[3] Barlevy, Gadi and Derek Neal. 2011. "Pay for Percentile." NBER working paper 17194.

[4] Bloom, Howard S. 2005. "Randomizing Groups to Evaluate Place-Based Programs." in *Learning more from social experiments: evolving analytic approaches*, ed. Howard S. Bloom, Russell Sage Foundation, New York.

[5] Cizek, Gregory J. 1999. *Cheating on Tests: How to do it, Detect it and Prevent it.* Mahwah, NJ: Lawrence Erlbaum.

[6] Cizek, Gregory J., Michael B. Bunch, and Heather Koons. 2004. "Standard Setting: A guide to establishing and evaluating performance standards on tests." *Educational Measurement: Issues and Practice*. 23.

[7] Cox, D. R., and N. Reid. 2000. *The Theory of the Design of Experiments*. Boca Raton, FL: Chapman and Hall.

[8] Fryer, Roland. 2010. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." NBER Working Paper No. 15898.

[9] Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2: 205-27.

[10] Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools;" *Journal of Economic Literature* 24: 1141-1177.

[11] Hanushek, Eric A. 2003, "The Failure of Input-Based Schooling Policies." *The Economic Journal* 113, F64-F98

[12] Heckman, James J, 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *The Quarterly Journal of Economics*, 115, No. 1, 45-97.

[13] Heckman, James J. and Sergio Urzua. 2010. "Comparing IV with structural models: What simple IV can and cannot identify." *Journal of econometrics*,156 iss:1. 27 -37

[14] Heckman, James J. and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *The Journal of Economic Perspectives*, 9, No. 2 , 85-110

[15] Hedges, L. V., Laine, R., and Greenwald, R. 1994. "Does Money Matter: A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Educational Researcher* 23: 5-14.

[16] Imbens, G., G. King, D. McKenzie, and G. Ridder 2009. On the Benefits of Stratification in Randomized Experiments Unpublished manuscript, Department of Economics, Harvard University.

[17] Jackson, C. Kirabo. 2010. "A Little Now for a Lot Later: A Look at a Texas Advanced Placement Incentive Program." *J. Human Resources* 45:591-639.

[18] Jacob, Brian and Steven Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics 118: 843-877.*

[19] Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn." *Review of Economic Statistics* 91: 437-456.

[20] Krueger, Alan B. 2003. "Economic Considerations and Class Size." *The Economic Journal* 113: F34-F63.

[21] Levitt, Steven D., John A. List, and Sally Sadoff. 2010. "The Effect of Performance-Based Incentives on Educational Achievement: Evidence from a Randomized Experiment." working paper, University of Chicago.

[22] Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Incentives in Developing Countries: Experimental Evidence from India." *Journal of Political Economy* 119: 39-77.

[23] Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92: 263-83.

[24] Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. "Teacher Pay for Performance Experimental Evidence from the Project on Incentives in Teaching." Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

[25] Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* 113: F3-F33.

[26] Todd, Petra E., and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *The American Economic Review* 96: 1384-1417.

[27] Todd, Petra E., and Kenneth I. Wolpin. 2012. "Estimating a Coordination Game Within the Classroom." Mimeo, University of Pennsylvania.

Appendix A

As discussed in Bloom (2005), carrying out power calculations before implementing an experiment requires some guesswork, because one needs an idea of the variance of the outcome and some estimate of the intraclass correlation. Once the experiment has been implemented, however, these preliminary estimates can be refined using the experimental data. As an easy way of understanding the power of an experimental design, Bloom (2005) introduces the notion of a minimum detectable effect size ($MDES$) for a given level of power (k), significance level (α), group size (n), number of groups (J), intraclass correlation (ρ) and for a proportion of subjects allocated to the treatment group (P). The $MDES$ is the program impact (divided by the standard deviation of the outcome for the target population) that can be detected with the specified parameters (power, significance level) under a particular experiment. The formula for the $MDES$ is

$$MDES = \frac{M_{J-2}}{J^{\frac{1}{2}}} \left(\rho + \frac{1-\rho}{n} \right)^{\frac{1}{2}} \left(\frac{1}{P(1-P} \right)^{\frac{1}{2}} \tag{4}$$

where M_{J-2} varies depending on whether a one-tail or two-tail t-test is conducted; $M_{J-2} = t_{1-k} + t_{\frac{\alpha}{2}}$ for a two-tailed test and $M_{J-2} = t_{1-k} + t_{\alpha}$ for a one-tail test. In the case of ALI, the average group size (n) is 200 students per school per grade. The number of groups is 28 for the control group and 20 for the treatment group, for a total (J) of 48 and a proportion (P) equal to 0.4167. For the calculations below, we require a power of 0.8 and assume an intraclass correlation (ρ) equal to 0.12, which is the estimated correlation using the full sample (all 88 schools) for the baseline 9th grade mathematics ENLACE (for each grade). With these parameters, the minimum detectable effect sizes ($MDES$) in a one-tailed test are 0.34, 0.26 and 0.22 for the three critical values α=.01, .05 and .10. Similarly, at those same critical values, the $MDES's$ are 0.37, 0.30 and 0.26 for a two tailed test.

As noted above, the required sample size can be substantially smaller if the analysis is conditioned on baseline school-level covariates. In the case of ALI, lagged school-level baseline student test scores are powerful predictors of current test scores and including them in a treatment impact regression substantially reduces the estimated intraclass correlation. Therefore, the $MDES$ reported above are conservative.

Appendix B

I. Calculation of the Student Bonus

1. *For 10th grade students*:

s = semester : s = 1,2.

c_{ks} = mathematics class designation k in semester s

$n_{\{ij\}}^{c_{ks}}$ = number of students in class k in semester s with test scores that place them in the ij^{th} cell of the 10th grade performance bonus schedule

$b_{\{ij\}}$ = bonus payment to a student in the ij^{th} cell of the 10th grade performance bonus schedule

$B_{\{ij\}}^{mn}$ = total bonus payment to a student with a test score in the $\{ij\}^{th}$ cell of the 10th grade performance bonus schedule who was in class m in semester one ($s = 1$) and class n in semester two ($s = 2$)

Bonus Formula

$$B_{\{ij\}}^{mn} = b_{\{ij\}} + \left(\sum_{\{ij\}} \frac{1}{2} \left[n_{\{ij\}}^{c_{m1}} b_{\{ij\}} + n_{\{ij\}}^{c_{n2}} b_{\{ij\}} \right] \right) \times .01$$

2. *For 11th grade students*:

s = semester : s = 3,4

c_{ks} = mathematics class designation k in semester s

$n_{\{ij\}}^{c_{ks}}$ = number of students in class k in semester s with test scores that place them in the ij^{th} cell of the 11th grade performance bonus schedule

$b_{\{ij\}}$ = bonus payment to a student in the ij^{th} cell of the 11th grade performance bonus schedule

$B_{\{ij\}}^{mn}$ = total bonus payment to a student with a test score in the $\{ij\}^{th}$ cell of the 11th grade performance bonus schedule who was in class m in semester one ($s = 3$) and class n in semester two ($s = 4$)

Bonus Formula

$$B_{\{ij\}}^{mn} = b_{\{ij\}} + \left(\sum_{\{ij\}} \frac{1}{2} \left[n_{\{ij\}}^{c_{m3}} b_{\{ij\}} + n_{\{ij\}}^{c_{n4}} b_{\{ij\}} \right] \right) \times .01$$

3. For 12th Grade Students:

B = bonus payment

Bonus Formula

$B = \$5,000$ if test score is "Proficient"

$\$10,000$ if test score is "Advanced"

II. Calculation of Mathematics Teacher Bonus

s = semester : s = 1,2 for 10th grade; s=3,4 for 11th grade; s=5,6 for 12th grade

$n^s_{\{ij\}k}$ = number of teacher $k's$ students in semester s with test scores that place them in the ij^{th} cell of the performance bonus schedule

$b^s_{\{ij\}}$ = bonus payment per student in the ij^{th} cell of the performance bonus schedule for semester s, where $b^1_{\{ij\}} = b^2_{\{ij\}}, b^3_{\{ij\}} = b^4_{\{ij\}}$ and $b^5_{\{ij\}} = b^6_{\{ij\}}$

f_k = teacher $k's$ Full-Time Equivalence (FTE) status = number of math classes taught over entire year (both semesters) divided by 10.[1]

M = total number of Mathematics teachers over the entire year regardless of the number of classes taught.

$F = \sum_{m=1}^{m=M} f_m$ = total number of full-time equivalent Mathematics teachers in the school;

B^1_k = bonus payment to teacher k for the performance of teacher $k's$ students

B^2_k = bonus payment to teacher k for the performance of students school-wide

B_k = total bonus payment to teacher k.

$F - f_k$ = the number of full-time equivalent teachers subtracting off teacher k's FTE status.

Bonus Formula

$$B^1_k = \frac{1}{2}\sum_{s=1}^{s=6}\sum_{\{ij\}} n^s_{\{ij\}k} b^s_{\{ij\}}$$

$$B^2_k = \left(\frac{1}{(F - f_k)} \sum_{\substack{m=1 \\ m\neq k}}^{M} B^1_m\right) \times f_k \times .25$$

$$B_k = B^1_k + B^2_k$$

III. Calculation of Non-Mathematics Teacher, Director and Associate Director Bonus

s = semester : s = 1,2 for 10th grade; s=3,4 for 11th grade; s=5,6 for 12th grade

f_k = non-mathematics teacher or administrator $k's$ Full-Time Equivalence (FTE) status

M = total number of Mathematics teachers over the entire year regardless of the number of classes taught.

$F = \sum_{m=1}^{m=M} f_m$ = total number of full-time equivalent Mathematics teachers in the school;

[1] For example, a teacher who taught 5 classes in the Fall term and 5 classes in the Spring term will have an FTE status of one ($f_k = 1$), while a teacher who taught 4 and 3 classes in the two terms will have an FTE status of .7 ($f_k = .7$). It is possible to have an FTE status that is greater than one.

B_m^1 = bonus payment to mathematics teacher m for the performance of teacher $m's$ students

B_k = total bonus payment to non-mathematics teacher or administrator k.

F = the number of full-time equivalent mathematics teachers

Bonus Formula

Director

$$B_k = \left(\frac{1}{F} \sum_{m=1}^{M} B_m^1 \right) \times f_k \times .50$$

Non-Mathematics Teacher and Associate Administrator

$$B_k = \left(\frac{1}{F} \sum_{m=1}^{M} B_m^1 \right) \times f_k \times .25$$

Appendix C

Average Treatment Effects (ATE) with and without Adjustments for Copiers: All Program Years[a,b,c]

Grade	Year One AY: 2008/2009			Year Two AY: 2009/2010			Year Three AY: 2010/2011		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
With Copying Adjustment[d]									
Tenth Grade									
	2008/2009			2009/2010			2010/2011		
ATE	15.3	1.20	31.4	27.2	-0.41	43.9	28.2	12.1	57.0
(s.e.)	(5.02)	(6.32)	(5.84)	(4.63)	(5.68)	(6.62)	(4.67)	(5.62)	(9.49)
Eleventh Grade									
	2007/2008			2008/2009			2009/2010		
ATE	9.21	-5.18	14.5	15.7	-1.40	33.3	14.2	-2.55	25.2
(s.e.)	(5.62)	(5.36)	(6.93)	(5.72)	(6.42)	(8.20)	(4.63)	(4.37)	(5.90)
Twelfth Grade									
	2006/2007			2007/2008			2008/2009		
ATE	10.9	7.13	31.5	12.3	-6.19	25.1	13.0	0.84	41.2
(s.e.)	(6.86)	(6.47)	(6.48)	(6.06)	(6.56)	(6.68)	(8.08)	(7.94)	(14.9)
No Copying Adjustment									
Tenth Grade									
	2008/2009			2009/2010			2010/2011		
ATE	18.5	1.11	32.3	32.4	0.31	54.7	41.5	15.9	83.4
(s.e.)	(5.02)	(5.35)	(6.18)	(5.24)	(5.74)	(11.1)	(6.25)	(6.16)	(16.9)
Eleventh Grade									
	2007/2008			2008/2009			2009/2010		
ATE	22.4	-2.98	27.8	55.5	6.17	67.4	51.3	-1.36	106.4
(s.e.)	(7.22)	(6.74)	(9.93)	(7.51)	(6.91)	(12.7)	(9.05)	(8.89)	(25.6)
Twelfth Grade									
	2006/2007			2007/2008			2008/2009		
ATE	9.73	4.73	29.3	36.0	-1.81	44.6	42.3	7.33	90.2
(s.e.)	(7.04)	(6.62)	(6.67)	(7.32)	(6.30)	(7.99)	(8.15)	(7.97)	(21.3)

a. Based on regressions that include school averages of 9th year ENLACE in each of the three grades in the program year in which the test was taken, the school average 12th grade ENLACE for the AY 2007/2008, region dummies and state dummies for states in which there was at least one school in all treatment groups. Standard errors account for clustering at the school level.
b. Sample based on students who took ALI test in each year of possible enrollment in program years.
c. There are 11,896 observations in the 2006/07 cohort, 11,385 in the 2007/08 cohort, 11,314 in the 2008/09 cohort, 13,778 in the 2009/10 cohort and 17,813 in the 2010/11 cohort.
d. Copiers assigned 25th percentile score of control group non-cheaters.

Table 1
Comparison of Treatment, Control and other Federal Non-ALI Schools (2007-08 AY)

	C[1]	T1[2]	T2[3]	T3[4]	Non-ALI	
Number of Schools	28	20	20	20	269[5]	408[6]
Blocking Variables						
Mean Number of Students	582 (0.77)	632 (0.55)	609 (0.72)	550 (0.63)	656 (0.12)	691 (0.02)
Mean Graduation Rate (Percent)	58.3 (0.74)	60.4 (0.53)	56.2 (0.51)	57.9 (0.07)	55.3 (0.24)	54.2 (0.15)
Non-Blocking Variables						
Pct. Oportunidades	40.3 (0.99)	39.5 (0.90)	40.6 (0.97)	40.1 (0.97)	37.6 (0.42)	18.9 (0.00)
Mean Class Size	35.8 (0.42)	41.0 (0.15)	39.0 (0.41)	35.7 (0.97)	34.7 (0.56)	42.1 (0.00)
Pct. Teachers with University Degree	82.3 (0.67)	79.4 (0.74)	81.7 (0.16)	84.8 (0.66)	81.5 (0.74)	81.0 (0.60)
Pct. New Directors	25.0 (0.72)	25.0 (1.00)	30.0 (0.71)	40.0 (0.29)	29.4 (0.62)	25.2 (0.98)
Mean Distance to Nearest Fed. Upper Sec. School (km)	32.9 (0.99)	32.8 (0.97)	31.4 (0.81)	32.4 (0.91)	23.9 (0.00)	15.9 (0.00)
Pct. DGETI	46.4 (0.92)	50.0 (0.81)	55.0 (0.57)	45.0 (0.92)	33.8 (0.20)	80.1 (0.00)
Pct. Region 1	35.7	35.0	50.0	50.0	30.8	36.0
Pct. Region 2	39.3	45.0	40.0	35.0	47.0	50.1
Pct. Region 3	17.9	10.0	5.0	10.0	15.4	10.4
Pct. Region 4	7.1	10.0	5.0	5.0	6.8	3.5
	(0.94)	(0.88)	(0.51)	(0.76)	(0.89)	(0.43)

[1] P-value for test C=T1=T2=T3 in parentheses.
[2] P-value for test C=T1 in parentheses.
[3] P-value for test C=T2 in parentheses.
[4] P-value for test C=T3 in parentheses.
[5] P-value for test C=Non-ALI schools in parentheses. Like the ALI schools, these are schools that have one session per day (morning).
[6] P-value for test C=Non-ALI schools in parentheses. These are schools that have two sessions per day (morning and afternoon). The figures pertain only to the morning session because data for afternoon sessions were often missing.

Table 2

Ninth Grade ENLACE: Treatment and Control Schools at Baseline

Variables	C[1]	T1[2]	T2[3]	T3[4]
9th Grade ENLACE Mean Test Score in Mathematics – Fall term enrollees[5]				
10th grade class	515.9	519.6	512.6	522.6
	(0.86)	(0.81)	(0.68)	(057)
11th grade class	516.0	516.6	517.4	524.7
	(0.91)	(0.96)	(0.86)	(0.47)
Pct. with ENLACE Score				
10th grade class	90.6	88.7	88.8	86.8
	(0.30)	(0.23)	(0.44)	(0.08)
11th grade class	78.3	74.0	75.2	75.3
	(0.62)	(0.25)	(0.37)	(0.39)

1. P-value for test C=T1=T2=T3 in parentheses; accounts for school-level clustering.
2. P-value for test C=T1 in parentheses; accounts for school-level clustering.
3. P-value for test C=T2 in parentheses; accounts for school-level clustering.
4. P-value for test C=T3 in parentheses; accounts for school-level clustering.
5. National mean is 500 and standard deviation 100.

Table 3
Schedule of Incentive Payments (Pesos) for Student Achievement

	End of Grade			
	Pre-Basic	Basic	Proficient	Advanced
Start of Grade				
10th Grade				
Pre-Basic	$0	$4000	$9000	$15000
Basic	$0	$2500	$7500	$13500
Proficient	$0	$0	$6000	$12000
Advanced	$0	$0	$4500	$10500
11th Grade				
Pre-Basic	$0	$4000	$9000	$15000
Basic	$0	$0	$7500	$13500
Proficient	$0	$0	$6000	$12000
Advanced	$0	$0	$4500	$10500
12th Grade				
Pre-Basic	$0	$0	$5000	$10000
Basic	$0	$0	$5000	$10000
Proficient	$0	$0	$5000	$10000
Advanced	$0	$0	$5000	$10000

Table 4
Schedule of Incentive Payments Per-Student for Mathematics Teachers

	End of Grade			
	Pre-Basic	Basic	Proficient	Advanced
Start of Grade				
10^{th} Grade				
Pre-Basic	0	$200	$450	$750
Basic	-$125	$125	$375	$675
Proficient	-$125	-$125	$300	$600
Advanced	-$125	-$125	$225	$525
11^{th} Grade				
Pre-Basic	0	$200	$450	$750
Basic	-$125	0	$375	$675
Proficient	-$125	-$125	$300	$600
Advanced	-$125	-$125	$225	$525
12^{th} Grade				
Pre-Basic	0	0	$250	$500
Basic	0	0	$250	$500
Proficient	0	0	$250	$500
Advanced	0	0	$250	$500

Table 5
Continuation Rates and ALI Test Completion Rates for 10th Grade, Year 1 Cohort

	Control Schools	Treatment 1 Schools[1]	Treatment 2 Schools[2]	Treatment 3 Schools[3]
Pct. Enrolled in Spring of Year 1 Given Enrollment in the Fall	90.1	88.0 (.384)	90.5 (.851)	86.8 (.203)
Pct. Taking ALI Exam in Year 1 Given Enrollment in Both Semesters	85.9	88.9 (.255)	85.6 (.926)	88.6 (.388)
Pct. Taking ALI Exam in Year 2 Given Test Taken in Year 1	.774	.819 (.053)	.769 (.820)	.808 (.172)
Pct. Taking ALI Exam in Year 3 Given Test Taken in Year 1	.678	.709 (.317)	.664 (.567)	.703 (.375)

1. P-value for test C=T1 in parentheses; accounts for school-level clustering.
2. P-value for test C=T2 in parentheses; accounts for school-level clustering.
3. P-value for test C=T3 in parentheses; accounts for school-level clustering.

Table 6

Percentage of Students with Non-Independent Test Scores by Year, Grade and Treatment

	Grade 10		Grade 11		Grade 12	
	Percentage Copiers	Percentage Cheaters	Percentage Copiers	Percentage Cheaters	Percentage Copiers	Percentage Cheaters
Year 1						
C	3.7	6.4	4.5	7.8	5.7	9.3
T1	5.1	9.1	10.9	14.9	5.2	8.4
T2	3.4	5.8	3.9	6.5	3.7	6.5
T3	3.7	6.7	10.1	14.9	2.7	4.7
Year 2						
C	3.5	6.1	3.6	6.2	2.4	4.5
T1	6.4	11.0	19.1	27.6	12.7	17.3
T2	4.3	7.4	6.2	9.8	3.4	5.5
T3	6.6	10.6	17.2	23.9	10.6	16.0
Year 3						
C	3.1	5.7	4.6	7.8	2.5	4.7
T1	8.1	13.2	19.8	28.2	17.5	24.7
T2	4.2	7.3	4.1	7.1	4.0	6.8
T3	10.3	16.2	23.8	31.3	15.4	21.3

Table 7

Average Treatment Effects (ATE) with and without Adjustments for Copiers: All Program Years[a,b,c]

Grade	Year One AY: 2008/2009			Year Two AY: 2009/2010			Year Three AY: 2010/2011		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
<u>With Copying Adjustment</u>									
Tenth Grade									
ATE	16.9	1.27	31.4	29.1	0.11	46.6	32.3	13.5	63.4
(s.e.)	(4.90)	(5.74)	(5.79)	(4.57)	(5.34)	(7.61)	(4.77)	(5.54)	(10.4)
P-value: TJ = T3	.010	<.001	-	.040	<.001	-	.002	<.001	-
Eleventh Grade									
ATE	13.6	-4.84	18.6	29.7	2.11	43.7	25.2	-2.00	42.1
(s.e.)	(5.40)	(5.50)	(7.39)	(4.89)	(6.05)	(8.33)	(4.24)	(4.31)	(5.64)
P-value: TJ = T3	.545	.004	-	.098	<.001	-	.011	<.001	-
Twelfth Grade									
ATE	9.63	4.71	28.8	21.9	-4.46	34.8	22.7	3.99	56.7
(s.e.)	(6.85)	(6.58)	(6.36)	(5.04)	(6.10)	(6.46)	(7.49)	(7.54)	(15.1)
P-value: TJ = T3	.010	<.001	-	.078	<.001	-	.015	<.001	-
<u>No Copying Adjustment</u>									
Tenth Grade									
ATE	18.5	1.11	32.3	32.4	0.31	54.7	41.5	15.9	83.4
(s.e.)	(5.02)	(5.35)	(6.18)	(5.24)	(5.74)	(11.1)	(6.25)	(6.16)	(16.9)
P-value: TJ = T3	.025	<.001	-	.073	<.001	-	.014	<.001	-
Eleventh Grade									
ATE	22.4	-2.98	27.8	55.5	6.17	67.4	51.3	-1.36	106.4
(s.e.)	(7.22)	(6.74)	(9.93)	(7.51)	(6.91)	(12.7)	(9.05)	(8.89)	(25.6)
P-value: TJ = T3	.639	.006	-	.403	<.001	-	.037	<.001	-
Twelfth Grade									
ATE	9.73	4.73	29.3	36.0	-1.81	44.6	42.3	7.33	90.2
(s.e.)	(7.04)	(6.62)	(6.67)	(7.32)	(6.30)	(7.99)	(8.15)	(7.98)	(21.3)
P-value: TJ = T3	.011	<.001	-	.400	.<.001	-	.022	<.001	-

a. Based on regressions that include school averages of 9th year ENLACE in each of the three grades in the program year in which the test was taken, the school average 12th grade ENLACE for the AY 2007/2008, region dummies and state dummies for states in which there was at least one school in all treatment groups. Standard errors account for clustering at the school level.

b. Sample based on students who took ALI test in each year of possible enrollment in program years.

c. There are 11,896 observations in the 2006/07 cohort, 11,385 in the 2007/08 cohort, 11,314 in the 2008/09 cohort, 13,778 in the 2009/10 cohort and 17,813 in the 2010/11 cohort. One T2 school missing pre-program 12th grade ENLACE scores.

Table 8
Average Treatment Effects by Gender and by 9th Grade ENLACE: 2008-09 Tenth Grade Cohort

	Tenth Grade (Year 1)			Eleventh Grade (Year 2)			Twelfth Grade (Year 3)		
	T1-C	T2-C	T3-C	T1-C	T2-C	T3-C	T1-C	T2-C	T3-C
Adjusted Score									
Gender									
Female	18.7	1.51	35.8	33.8	4.71	51.0	28.8	6.72	63.9
	(5.65)	(6.39)	(5.30)	(5.62)	(6.40)	(7.43)	(7.85)	(7.57)	(15.8)
Male	15.0	1.32	33.0	25.3	-0.32	45.5	14.7	-1.10	63.7
	(5.91)	(6.42)	(7.48)	(5.79)	(6.64)	(9.98)	(7.84)	(9.03)	(14.9)
9th Grade ENLACE									
Pre-Basic	15.0	1.95	26.8	24.4	2.11	33.4	23.6	4.75	50.7
	(4.07)	(4.49)	(4.84)	(3.59)	(4.79)	(5.98)	(6.28)	(6.32)	(12.7)
Basic	18.2	-1.70	30.8	35.3	-0.15	48.9	22.5	2.72	57.4
	(5.92)	(7.43)	(7.71)	(5.95)	(7.32)	(9.54)	(8.87)	(8.76)	(16.6)
Proficient or Advanced	28.0	1.19	45.3	47.3	-2.12	58.1	45.6	17.9	70.2
	(12.5)	(16.1)	(17.5)	(13.5)	(16.1)	(19.8)	(16.1)	(17.7)	(23.7)
Unadjusted Score									
Gender									
Female	19.9	0.62	37.2	54.7	5.91	77.6	44.8	9.04	101
	(5.87)	(5.73)	(5.76)	(7.54)	(6.54)	(9.94)	(8.93)	(8.32)	(22.9)
Male	17.1	2.09	34.6	55.2	5.90	76.2	37.2	2.21	99.4
	(6.01)	(6.42)	(7.62)	(8.11)	(8.33)	(12.8)	(7.75)	(9.76)	(18.7)
9th Grade ENLACE									
Pre-Basic	16.4	1.44	28.8	45.4	2.94	65.3	43.9	6.89	97.0
	(4.46)	(4.87)	(6.26)	(7.21)	(5.96)	(11.5)	(7.59)	(7.92)	(22.4)
Basic	21.0	-0.67	31.8	64.0	5.06	69.7	39.9	7.27	80.6
	(6.10)	(6.89)	(7.90)	(8.85)	(8.38)	(14.8)	(9.32)	(8.79)	(21.2)
Proficient or Advanced	27.8	-0.26	44.3	73.6	5.86	72.4	61.4	25.2	86.3
	(12.4)	(15.9)	(17.5)	(16.0)	(17.0)	(22.1)	(17.6)	(18.5)	(26.7)

a. There are 5388 males, 5665 females, 4926 observations in the pre-basic category, 4294 observations in the basic category and 1223 observations in the proficient or advanced category. Standard errors, in parentheses, account for clustering at the school level.

Table 9

Student and Teacher Effort Measures by for Controls and Treatment/Control Difference: Year 3

	C			T1 - C			T2 - C			T3 - C		
Grade	10	11	12	10	11	12	10	11	12	10	11	12
Student:												
Avg, hrs/wk study math	4.68	4.45	4.53	.199	.408	.385	-.138	-.070	-.097	.397	.301	.370
				(.095)	(.135)	(.124)	(.091)	(.182)	(.165)	(.112)	(.135)	(.127)
Avg, hrs/wk study non-math subjects	5.56	5.48	5.32	.109	.189	.250	-.161	-.134	-.040	.152	.074	.168
				(.122)	(.122)	(.156)	(.129)	(.168)	(.153)	(.122)	(.134)	(.127)
Frac. pay attention >75% of time	.473	.479	.481	.070	.048	.042	.015	.007	-.006	.101	.070	.050
				(.022)	(.021)	(.024)	(.028)	(.030)	(.026)	(.028)	(.023)	(.032)
Frac. never or almost never text while doing homework	.423	.429	.415	.109	.093	.056	.023	.004	-.007	.126	.097	.061
				(.023)	(.028)	(.027)	(..026)	(.028)	(.028)	(.024)	(.022)	(.022)
Frac. never or almost never watch TV while doing homework	.493	.517	.498	.077	.075	.066	-.021	-.010	-.010	.088	.093	.060
				(.028)	(.018)	(.024)	(.025)	(.022)	(.020)	(.026)	(.022)	(.027)
Frac. Gave Help to Classmates	.599	.608	.643	.055	.058	.026	-.017	-.014	-.041	.086	.087	.026
				(.020)	(.022)	(.023)	(.020)	(.019)	(.028)	(.020)	(.022)	(.028)
Frac. Report Putting Much Effort	.466	.489	.486	.077	.090	.087	-.039	-.029	-.017	.114	.093	.092
				(.022)	(.026)	(.028)	(.021)	(.030)	(.025)	(.022)	(.021)	(.037)
Teacher:												
Frac. prepared students for ALI test	.167	.260	241	.202	.181	.211	.182	.155	.111	.412	.256	.176
				(.103)	(.121)	(.107)	(.091)	(.106)	(.114)	(.106)	(.110)	(.098)
Frac. helped students outside of class to prepare for ALI test	.241	.220	.204	.338	.339	.453	.341	.390	.391	.435	.554	.482
				(.104)	(.126)	(.102)	(.103)	(.111)	(.122)	(.098)	(.092)	(.103)

a. Standard errors, in parentheses, corrected for clustering at school level

Table 10
Pct. Receiving Payment and Incentive Payment Cost (Pesos) – Year Two

	Treatment 3	Treatment 1	Treatment 2
Pct. of Students Receiving Payment			
Grade 10			
For Own Performance	64.6	58.8	
For Class Performance	100.0	-	
Grade 11			
For Own Performance	41.3	38.8	
For Class Performance	99.4	-	
Grade 12			
For Own Performance	17.3	15.3	
Mean Student Payment:			
Grade 10			
For Own Performance	2,991	2,515	
For Class Performance	1,108	-	
Total	4,099	2,515	
Grade 11			
For Own Performance	2,679	2,541	
For Class Performance	861	-	
Total	3,540	2,541	
Grade 12			
For Own Performance	991	915	
Pct. of Teachers Receiving Payment			
For Own Performance	97.2		93.5
For Class Performance	100.0		-
Mean Math Teacher Payment (FTE):			
For Own Performance	15,330		6,332
For Other Teacher Performance	3,779		-
Total	19,109		6,332
Mean Non-Math (NM) Teacher and Assistant Director (AD) Payments			
Payment per FTE	3,872		-
Mean Director Payments:			
Payment per Director	7,744		-
Incentive Payment Cost Per-Student	**3,303**	**2,080**	**43**
Amount controls would receive	**1,643**	**1,163**	**44**
Pct. of total	**49.7**	**55.9**	**100**