American Economic Review 2009, 99:5, 1979–2011
http://www.aeaweb.org/articles.php?doi=10.1257/aer.99.5.1979

Performance Pay and Teachers' Effort, Productivity, and Grading Ethics

By VICTOR LAVY*

This paper presents evidence about the effect of individual monetary incentives on English and math teachers in Israel. Teachers were rewarded with cash bonuses for improving their students' performance in high-school matriculation exams. The main identification strategy is based on measurement error in the assignment to treatment variable that produced a randomized treatment sample. The incentives led to significant improvements in test taking rates, conditional pass rates, and mean test scores. Improvements were mediated through changes in teaching methods, enhanced after-school teaching, and increased responsiveness to students' needs. No evidence was found of manipulation of test scores by teachers. (*JEL* I21, J31, J45)

Performance-related pay for teachers is being introduced in many countries, amidst much controversy and opposition from teachers and unions alike.[1] The rationale for these programs is the notion that incentive pay may motivate teachers to improve their performance. However, there is little evidence of the effect of changes in teachers' incentives in schools. In this paper, I present evidence from an experimental program that offered teachers bonus payments on the basis of the performance of their classes in high-school matriculation exams in English and mathematics. The bonus program was structured as a rank-order tournament among teachers, separately by subject. Thus, teachers were rewarded on the basis of their performance relative to other teachers of the same subjects. Two measurements of student achievements were used as indicators of teachers' performance: the pass rate and the average score on each matriculation exam. The total amount awarded in each tournament was predetermined, and individual awards were determined on the basis of rank and an award scale. The main interest in this experiment relates to the effect of the program on teachers' pedagogy and effort, on teacher's productivity as measured by students' achievements, and on teachers' grading ethics.

* Department of Economics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel, NBER, and Department of Economics, Royal Holloway, University of London (e-mail: msvictor@huji.ac.il). Special thanks go to Alex Levkov, Yannay Spitzer, Roy Mill, and Katherine Eyal for outstanding research assistance. I also thank Josh Angrist, Abhijit Banerjee, Eric Battistin, Esther Duflo, Caroline M. Hoxby, Andrea Ichino, Hessel Oosterbeek, Yona Rubinstein, and seminar participants at EUI, MIT, Hebrew University, Princeton University, Tel Aviv University, The Tinbergen Institute, and the NBER Labor Studies Summer Institute for helpful discussions and comments. Finally, I thank the editor and referees for helpful comments. The 2001 Teachers' Incentive Program was funded by the Israel Ministry of Education and administered by the division for secondary schools. I also acknowledge funding from the Falk Institute for Economic Research in Israel. The views expressed in this paper are those of the author alone and have not been endorsed by the program sponsors or funding agency. This is a substantially revised version of NBER Working Paper 10622.

[1] Examples include performance-pay plans in Denver (2006) and in Houston (2006). Earlier examples are programs in Dade County, FL, and in Dallas, TX, in the mid-1990s: statewide programs in Iowa, Arizona, and California in 2002; and programs in Cincinnati, Philadelphia, and Coventry (Rhode Island). See David J. Wakelyn (1996); Nancy Mitchell, "Denver Teachers Opt for Merit Pay," *Rocky Mountain News*, December 29, 2005; Connie Sadowski (2006); and Lavy (2007) for discussion of some of these programs.

Although the program was designed as an experiment, schools were not assigned to it at random. Nevertheless, the design of the program enables the implementation of a quasi-randomized trial identification strategy based on two features of the program: assignment of schools to the program based on a threshold function of an observable, and a measurement error in this variable. Schools were included in the program if their 1999 matriculation rate was equal to or lower than a critical value (45 percent). This rule may be described as $T = 1\{S <= 45\}$, where T is an indicator of assignment to treatment and S is the assignment variable. However, there was a measurement error in S; thus, $S = S^* + \varepsilon$, where S^* is the true rate and ε is the measurement error. The administrators of the program, unaware that the assignment variable used was measured erroneously, assigned some schools to the program mistakenly. As I show below, ε appears to be essentially random and unrelated to the potential outcome. Therefore, for the group of schools around the threshold, the assignment of the treatment was random given their S^*. Therefore, by controlling for S^*, potentially in a fully nonparametric way, there was in fact conditional random assignment. In implementing this identification strategy, I also use available panel data (before and after the program) that allow an estimation of difference-in-differences estimates in this natural experiment setting.

The primary method of identification strategy I use in this paper is the one described above, but I also present results from two different approaches. The first of these alternatives is based on the notion of a regression discontinuity (RD) design, i.e., that the likelihood of an S value slightly above or below the threshold value of the assignment variable is largely random. If this is true, then treated and untreated schools in a narrow band around the threshold may be indistinguishable in their potential outcomes. However, a weaker assumption is based on controlling for parametric functions of S. In other words, conditional on S, we expect no variation in T. I exploit this sharp discontinuity feature in the assignment mechanism to estimate the effect of the teachers' incentive program. Here, as in the measurement-error method, I exploit the panel nature of the data and embed the RD design in a difference-in-differences estimation. The second alternative is based on comparing all treated schools to *all* eligible schools that were not chosen to participate in the program and using panel data to control for school fixed effects. This alternative also reduces the risk of omitted confounding students' effects by using the multidimension lagged student outcomes as controls. In principal, the estimates based on these two alternative approaches may be potentially biased but, in actuality, the RD method yields similar evidence to that obtained using the measurement-error randomized trial approach.

Section I of this paper provides background information about the Israeli school system, describes the teachers' incentive program, and discusses the theoretical context of pay-for-performance programs. Section II discusses identification, estimation, and results. Sections III and IV present evidence of the effect of incentives on teachers' effort, pedagogy, and grading ethics. Section V describes the broader context and generalizability of the Israeli experiment. Section VI concludes.

The results suggest that teachers' incentives increase student achievements by increasing the test taking rate as well as the conditional pass rate and test scores in math and English exams. The improvement in these conditional outcomes, which are estimated based on tests and grading external to schools, accounts for more than half of the increase in the unconditional outcomes in math and somewhat less in English. These improvements appear to result from changes in teaching methods, after-school teaching, and increased responsiveness to students' needs, and not from artificial inflation or manipulation in test scores. The evidence that incentives induced improved effort and pedagogy is important in the context of the recent concern that incentives may have unintended effects, such as "teaching to the test" or cheating and manipulation of test scores, and that they do not generate real learning.

I. Tournaments as a Performance Incentive

Formal economic theory usually justifies incentives to individuals as a motivation for efficient work. The underlying assumption is that individuals respond to contracts that reward performance. However, only a small proportion of jobs in the private sector base remuneration on explicit contracts that reward individual performance. The primary constraint in individual incentives is that their provision inflicts additional risks on employees, for which employers incur a cost in the form of higher wages. A second constraint is the incompleteness of contracts, which may lead to dysfunctional behavioral responses in which workers emphasize only those aspects of performance that are rewarded. These constraints may explain why private firms reward workers more through promotions and group-based merit systems than through individual merit rewards (Canice Prendergast 1999).

In education, too, group incentives are more prevalent than individual incentive schemes. The explanation for this pattern, it is argued, lies in the inherent nature of the educational process. Education involves teamwork, the efforts and attitudes of fellow teachers, multiple stakeholders, and complex jobs that involve multitasking. In such a working environment, it is difficult to measure the contribution of any given individual. The group (of teachers, in this case) is often better informed than the employer about its constituent individuals and their respective contributions, enabling it to monitor its members and encourage them to exert more effort or exhibit other appropriate behavior. It is also argued that individuals who have a common goal are more likely to help each other and make more strenuous efforts when a member of the group is absent. On the other hand, standard free-rider arguments cast serious doubt on whether group-based plans provide a sufficiently powerful incentive, especially when the group is quite large (Bent Holmström and Paul Milgrom 1991).

Tournaments as an incentive scheme were suggested initially as appropriate in situations where individuals exert effort in order to get promoted to a better paid position, where the reward associated with that position is fixed, and where there is competition among individuals for these positions (Edward Lazear and Sherwin Rosen 1982; Jerry Green and Nancy L. Stokey 1983). The only question that matters in winning such tournaments is how well one does relative to others—not the absolute level of performance. Although promotion is not an important career feature among teachers, emphasis on relative rather then absolute performance measures is relevant for a teacher-incentive scheme for two reasons. First, awards based on relative performance and a fixed set of rewards would stay within budget. Second, in a situation were there are no obvious standards that may be used as a basis for absolute performance, relying on how well teachers do relative to others seem a preferred alternative. Therefore, we used the structure of a rank-order tournament for the teacher-incentive experiment described below.

A. *Secondary Schooling in Israel*

High school students in Israel are tested in a series of Bagrut (matriculation) examinations, a set of national exams in core and elective subjects that begins in tenth grade, continues in eleventh grade, and concludes in twelfth grade, when most of the tests are taken. Pupils choose to be tested at various levels in each subject, each test awarding from one to five credit units (hereafter, credits) per subject.[2] The final matriculation score in a given subject is the mean of two interme-

[2] Many countries (Germany, France, Italy) have similar high school matriculation systems. The New York State Regents examinations and the Massachusetts Comprehensive Assessment System are similar. The curriculum and the external exam system of the Bagrut study program in Israel are also very similar to the system of Advanced Placement courses in the United States.

diate scores. The first is based on the score in the national exams that are "external" to the school because they are written, administered, supervised, and graded by an independent agency. The scoring process for these exams is anonymous; the external examiner is not told the student's name, school, or teacher. Exams are held in June and January, and all pupils are tested in a given subject on the same date. The national exams are graded centrally by two independent external examiners and the final score is the average of the two. The second intermediate score is based on a school-level ("internal") exam that mimics the national exam in material and format but is scored by the student's own teacher.

Some subjects are mandatory and many must be taken at the level of three credits at least. Tests that award more credits are more difficult. English and math are among the core compulsory subjects and must be studied at one of three levels: basic (three credits), intermediate (four credits) and advanced (five credits). A minimum of 20 credits is required to qualify for a matriculation certificate. About 45 percent of high school seniors received matriculation certificates in 1999 and 2000, i.e., passed enough exams to be awarded 20 credits and satisfied distributional requirement by the time they graduated from high school or shortly thereafter (Israel Ministry of Education 2001). The high school matriculation certificate in Israel is a prerequisite for university admission and is one of the most economically important education milestones.

B. *The Israeli Teacher-Incentive Experiment*

In early December 2000, the Ministry of Education unveiled a new teachers' bonus experiment in 49 Israeli high schools.[3] The main feature of the program was an individual performance bonus paid to teachers on the basis of their own students' achievements. The experiment included all English, Hebrew, Arabic, and mathematics teachers who taught classes in grades 10 through 12 in advance of matriculation exams in these subjects in June 2001. In December 2000, jointly with the Ministry, I conducted an orientation activity for principals and administrators of the 49 schools. The program was described to them as a voluntary three-year experiment.[4] All the principals reacted very enthusiastically to the details of the program except for one, who decided not to participate.

Schools were also allowed to replace the language (Hebrew and Arabic) teachers with teachers of other core matriculation subjects (Bible, literature, or civics). Therefore, school participation in Hebrew and Arabic was not compulsory but a choice of the school. This choice may have been correlated with potential outcome, i.e., the probability of success of teachers in the tournament, resulting in an endogenous participation in the program in that subject. Therefore, the evaluation may include only English and math teachers.

Each of the four tournaments (English, Hebrew and Arabic, math, and other subjects) included teachers of classes in grades 10–12 that were about to take a matriculation exam in one of these subjects in June 2001. Each teacher entered the tournament as many times as the number of classes he/she taught and was ranked each time on the basis of the mean performance of each of his/her classes. The ranking was based on the difference between the actual outcome and a value predicted on the basis of a regression that controlled for the students' socioeconomic characteristics, the level of their study program in the relevant subject (basic, intermediate, and advanced), grade (10th, 11th and 12th), grade size, and a fixed school-level effect.[5] The school fixed effects

[3] Another program, based on students' bonuses, was conducted simultaneously in a different set of schools. There is no overlap of schools in these different incentive programs, either in the treatment or control groups of this study.

[4] Due to the change in government in March 2001 and the budget cuts that followed, the Ministry of Education announced in the summer of 2001 that the experiment would not continue as planned for a second and third year.

[5] Note that the regression used for prediction did not include lagged scores, in order that teachers would have no incentive to play the system, for example by encouraging students not to do their best in earlier exams that do not count

imply that the predicted values were based on within school variation among teachers and the teachers were told explicitly that they were to be compared to other teachers of the same subject in the same school. Separate regressions were used to compute the predicted pass rate and mean score, and each teacher was ranked twice—once for each outcome—using the size of the residual from the regressions. The school submitted student enrollment lists that were itemized by grades, subjects, and teachers. The reference population was those enrolled on January 1, 2001, the starting date of the program. All students who appeared on these lists but did not take the exam (irrespective of the reason) were assigned an exam score of zero.

All teachers whose students' mean residual (actual outcome less predicted outcome) was positive in both outcomes were divided into four ranking groups, from first place to fourth. Points were accumulated according to ranking: 16 points for first place, 12 for second, 8 for third, and 4 for fourth. The program administrators gave more weight to the pass rate outcome, awarding a 25 percent increase in points for each ranking (20, 15, 10, and 5, respectively). The total points in the two rankings were used to rank teachers in the tournament and to determine winners and awards, as follows: 30–36 points—$7,500; 21–29 points—$5,750; 10–20 points—$3,500; and 9 points—$1,750. These awards are significant relative to the mean gross annual income of high-school teachers ($30,000) and the fact that a teacher could win several awards in one tournament if he or she prepared more than one class for a matriculation exam.[6] Since the program was revealed to teachers only in the middle of the year, it is unlikely that there was a teachers' selection based on the expectation of an increased income or that teachers could have manipulated their class composition.

Three formal rules guided the assignment of schools to the program: only comprehensive high schools (comprising grades 7–12) were eligible, the schools must have a recent history of relatively poor performance in the mathematics or English matriculation exams,[7] and the most recent school-level matriculation rate must be equal to or lower than the national mean (45 percent). A total of 106 schools met the first two criteria but 7 of them were disqualified because they were already part of other remedial education programs; therefore there were 99 eligible schools of which 49 met the third criteria. However, as noted above, as one school declined to participate in the program, the actual number of participants was 48 schools (treated sample).[8]

The program included 629 teachers, of whom 207 competed in English, 237 in mathematics, 148 in Hebrew or Arabic, and 37 in other subjects that schools preferred over Hebrew. Awards were granted to 302 teachers—94 English teachers, 124 math teachers, 67 Hebrew and Arabic teachers, and 17 among the other subjects. Three English teachers won two awards each, 12 math teachers won two awards each, and one Hebrew teacher won two first-place awards totaling $15,000.

We conducted a follow-up survey during the summer vacation at the end of the school year,. and 74 percent of teachers in the program were interviewed. Very few of the intended interviewees were not interviewed. Failure to be interviewed was mostly due to incorrect telephone numbers or teachers who could not be reached by telephone after several attempts. The survey results show that 92 percent of the teachers knew about the program, 80 percent had been briefed

in the tournament. This feature would have been important had the program continued.

[6] For more details, see Israel Ministry of Education, High School Division, "Individual Teacher Bonuses Based on Student Performance: Pilot Program," December 2000, Jerusalem (Hebrew).

[7] Performance was measured in terms of the average pass rate in the mathematics and English matriculation tests during the past four years (1996–1999). If any of these rates were lower than 70 percent in two or more occurrences, the school's performance was considered poor. English and math were chosen because they have the highest failure rate among matriculation subjects.

[8] Since a relatively large number of religious and Arab schools were included in the eligible sample (higher than their proportion in the sample), the matriculation threshold for these schools was set at 43 percent.

about its details—almost all by their principals and the program coordinator—and 75 percent thought that the information was complete and satisfactory. Almost 70 percent of the teachers were familiar with the award criteria and about 60 percent of them thought they would be among the award winners. Only 30 percent did not believe they would win; the rest were certain about their chances. Two-thirds of the teachers thought that the incentive program would lead to an improvement in student achievement.

C. *The Data*

The data used in this study pertain to the school year preceding the program, September 1999–June 2000, and the school year in which the experiment was conducted, September 2000–June 2001. The school data provide information on the ethnic (Jewish or Arab) nature of each school, the religious orientation (secular or religious) of the Jewish schools, and each school's matriculation rate in the years 1999–2001. The micro student files included the full academic records of each student for the Bagrut exams during high school (grades 10–12) and student characteristics (gender, parental schooling, family size, immigration status—students who recently immigrated). The information for each Bagrut exam included its date, subject, applicable credits, and score. The base sample is all 12th grade students in 2000 and 2001. A very small proportion of the students completed their math and/or English requirement by the end of 11th grade and therefore was not subject to the intervention. These students were therefore excluded from the analysis and thus the math and English samples are not identical.

I defined three outcomes for each subject based on the summer (June 2001) period of exams: an indicator of taking the test in a given subject, an indicator of passing the exam (a score $=>$ 55) and the actual test score (from 1 to 100). The latter two were the criteria used to rank teachers to determine award winners. Table 1 presents descriptive statistics for the 2000 and 2001 cohorts of high school seniors for two samples: the 48 schools included in the program and all other 50 eligible high schools.[9] The standard errors reported in the table are adjusted for clustering within schools. Panel A shows that the treated sample included relatively more Arab schools, and this difference is significantly different from zero.[10] Panel A also shows the large gap in the lagged matriculation rate between the treated and nontreated schools. In 1999 this gap was 25.1 percent and in 2000 it was 23.0 percent. Panels B and C reveal that in the treatment and pretreatment years, the means of students' characteristics in treated schools differed from the corresponding means in all other eligible schools. For example, parental schooling is higher by more than two years in the untreated eligible schools. Large differences between the two sets of schools were also observed in the means of lagged students' outcomes. By implication, the sample of the 48 treated schools is not representative of the sample of all eligible schools. Therefore, a simple comparison between the treated and untreated eligible schools cannot be the basis for identification.

[9] The school that declined to participate in the program is excluded from the treated sample and from the eligible sample because it did not provide necessary data, and therefore the Ministry excluded it from data files provided for this study. Dropping this school from the analysis does not affect the main results of the paper because, based on its correct and erroneous 1999 matriculation rate, it would have not been included anyway in the randomized treatment (RT) or RD samples.

[10] These school characteristics are time invariant but there are minor differences in their means in 2000 and 2001 because they are computed from the student samples which are different in the two years.

TABLE 1—DESCRIPTIVE STATISTICS: TREATED SCHOOLS VERSUS ALL OTHER ELIGIBLE SCHOOLS

	2000			2001		
	Treated schools (1)	Nontreated schools (2)	Difference (3)	Treated schools (4)	Nontreated schools (5)	Difference (6)
Panel A. School characteristics						
Religious school[a]	0.199	0.269	−0.070 (0.084)	0.182	0.258	−0.076 (0.080)
Arab school[b]	0.260	0.099	0.161 (0.081)	0.284	0.107	0.176 (0.087)
Lagged Bagrut rate	0.369	0.620	−0.251 (0.027)	0.377	0.607	−0.230 (0.035)
Panel B. Student background						
Father education	9.062	11.386	−2.324 (0.639)	9.029	11.357	−2.329 (0.582)
Mother education	8.817	11.486	−2.669 (0.709)	8.551	10.846	−2.295 (0.751)
Number of siblings	3.463	2.580	0.883 (0.422)	3.472	2.481	0.991 (0.425)
Gender (male = 1)	0.495	0.466	0.028 (0.032)	0.508	0.492	0.016 (0.029)
Immigrant	0.031	0.027	0.004 (0.015)	0.022	0.010	0.012 (0.009)
Asia-Africa ethnicity	0.190	0.208	−0.018 (0.031)	0.170	0.190	−0.020 (0.030)
Panel C. Student lagged outcomes						
Math credits gained	0.290	0.499	−0.209 (0.132)	0.320	0.571	−0.251 (0.130)
English credits gained	0.127	0.194	−0.067 (0.047)	0.116	0.183	−0.067 (0.050)
Total credits attempted	4.292	5.283	−0.991 (0.320)	4.502	5.464	−0.962 (0.341)
Total credits gained	3.388	4.591	−1.203 (0.301)	3.633	4.773	−1.140 (0.303)
Average score	56.580	69.555	−12.974 (2.296)	58.381	69.699	−11.318 (2.010)
Observations	6,250	5,931	12,181	6,084	5,820	11,904
Number of schools	48	50	98	48	50	98

Note: Standard errors in parentheses are adjusted for school-level clustering.

[a] The schools status of nationality and religiosity does not change. Any change in the means across years reflects relative changes in the number of students in a cohort.

[b] This table is based on the math sample.

II. Identification, Estimation, and Results

A. *Natural Experiment Due to Random Measurement Error in the Assignment Variable*

The program rules limited assignment to schools with a 1999 matriculation rate equal to or lower than 45 percent (43 percent for religious and Arab schools). However, the matriculation rate used for assignment was an inaccurate measure of this variable. The data given to administrators were culled from a preliminary and incomplete file of matriculation status. For many

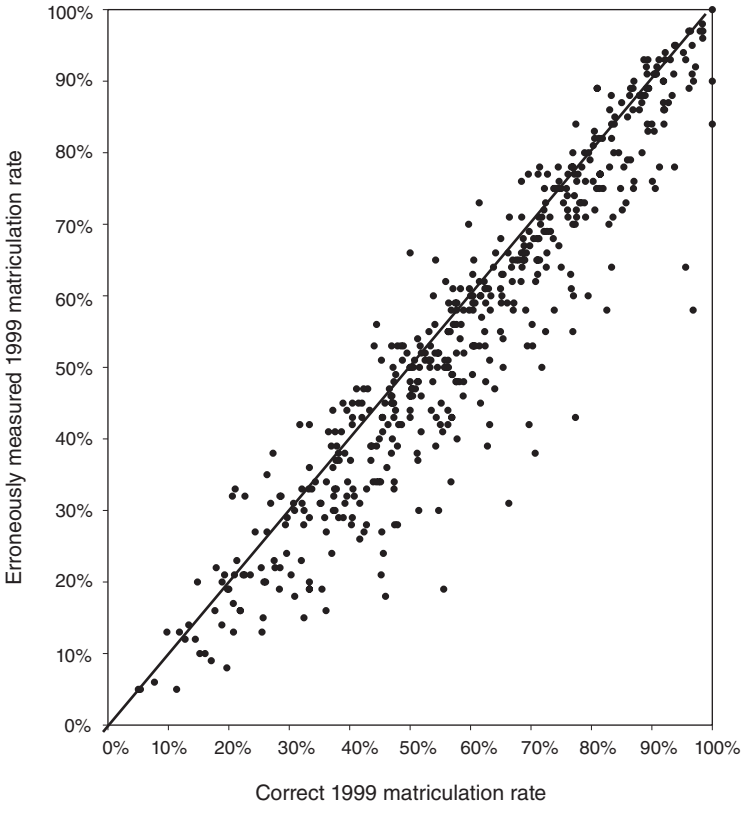

FIGURE 1. THE RELATIONSHIP BETWEEN THE CORRECT AND THE ERRONEOUSLY MEASURED 1999 MATRICULATION RATE

(Sample = 507 schools)

students, matriculation status was erroneous, as it was based on missing or incorrect information. The Ministry later corrected this preliminary file, as it does every year.[11] As a result, the matriculation rates used for assignment to the program were inaccurate in a majority of schools. This measurement error could be useful for identification of the program effect. In particular, conditional on the true matriculation rate, program status may be virtually randomly assigned due to mistakes in the preliminary file.

Figure 1 presents the relationship between the correct matriculation rates and those erroneously measured for a sample of 507 high schools in Israel in 1999.[12] Most (80 percent) of the measurement errors were negative, 17 percent were positive, and the rest were free of error. The deviations from the 45-degree line do not seem to correlate with the correct matriculation rate. This may be seen more clearly in Figure 2, which demonstrates that the measurement error and the matriculation rate do not covary; their correlation coefficient is very low, at –0.084, with a p-value that differs from zero (0.059). However, if a few extreme values (five schools) are

[11] There are many requirements to complete the matriculation process that tend to vary by school type and level of proficiency in each subject. The verification of information between administration and schools is a lengthy process. The first version of the matriculation data becomes available in October and is finalized in December.

[12] The sample was limited to schools with positive (> 5 percent) true matriculation rates.

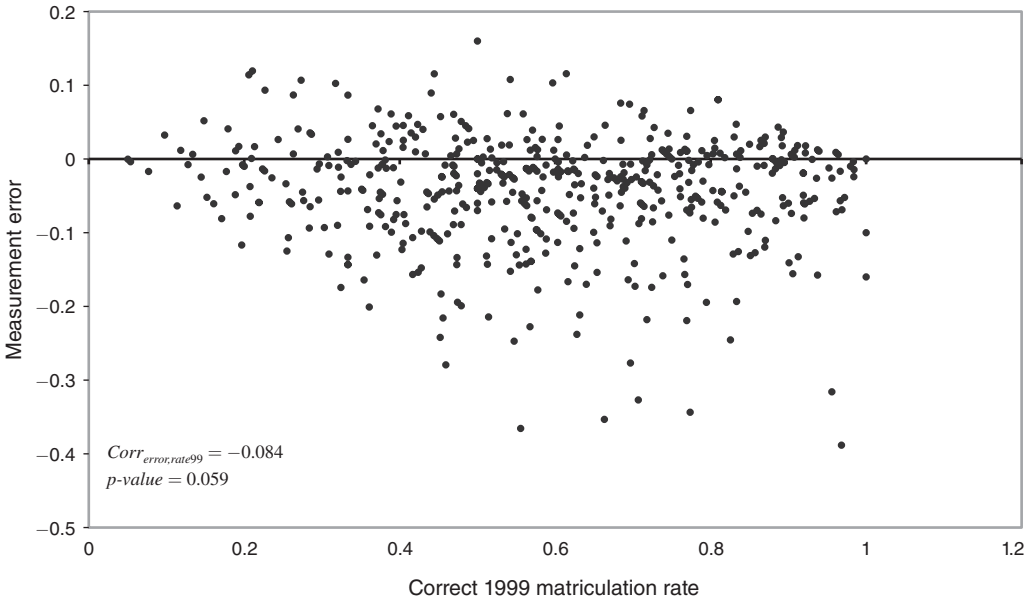

FIGURE 2. THE CORRECT 1999 MATRICULATION RATE VERSUS THE MEASUREMENT ERROR
(Sample = 507 schools)

excluded, the correlation coefficient is effectively zero. Although the figure possibly suggests that the variance of the measurement error is lower at low matriculation rates, this is most likely due to the floor effect that bounds the size of the negative errors: the lower the matriculation rate, the lower the absolute maximum size of the negative errors.[13]

A further check on the random nature of the measurement error may be based on its statistical association with other student or school characteristics. Table 2 presents the estimated coefficients from regressions of the measurement error on student characteristics, lagged students' outcomes, and school characteristics of the 2001 high school seniors. Each entry in the table is based on a separate regression which was run with school-level means of all variables, separately for the whole sample and for the eligible sample. These estimates are presented in columns 1 and 2, respectively.

There are 12 estimates for each sample, and only a few are significantly different from zero: three in the full sample estimates and two in the eligible sample. Furthermore, the variables that are significant are different across samples, suggesting that these are transitory and random differences. For example, in the eligible sample there is an imbalance in the proportion of immigrant students (the estimate is −0.364, s.e. 0.173) but in the sample of all schools this control-treatment difference is positive and practically zero (the estimate is 0.068, s.e. 0.046). Based on the evidence presented in Figures 1 and 2 and Table 2, it may be concluded that there is no evidence of a significant association between the measurement error in 1999 and the observable

[13] Similar results are observed when the sample is limited to schools with a matriculation rate higher than 40 percent. In this sample, the problem of the bound imposed on the size of the measurement error at schools with low matriculation rates is eliminated. I also examined a sample that was limited to the eligible schools (97 schools and not 98, because one of the eligible schools was missing the true 1999 matriculation rate). The results, which are presented in Figures A1 and A2 in the Web Appendix, available at http://www.aeaweb.org/articles.php?doi=10.1257/aer.99.5.1979, are identical to those in Figures 1 and 2.

TABLE 2—ESTIMATES FROM REGRESSIONS OF THE 1999 MEASUREMENT ERROR IN THE SCHOOL MATRICULATION RATE ON THE 2001 STUDENT AND SCHOOL CHARACTERISTICS

	All schools (1)	Eligible schools (2)	RT schools (3)
Panel A. School characteristics			
Religious schools	−0.008	−0.026	−0.065
	(0.007)	(0.017)	(0.031)
Arab school	−0.022	−0.013	0.038
	(0.009)	(0.021)	(0.058)
Panel B. Student background			
Father education	0.001	0.001	−0.011
	(0.001)	(0.003)	(0.007)
Mother education	0.000	−0.002	−0.008
	(0.001)	(0.002)	(0.006)
Number of sibblings	−0.008	−0.003	−0.007
	(0.002)	(0.005)	(0.014)
Gender (male = 1)	0.023	−0.001	−0.087
	(0.013)	(0.035)	(0.060)
Immigrant	0.068	−0.364	−0.674
	(0.046)	(0.173)	(0.247)
Panel C. Student lagged outcomes			
Math credits gained	0.010	0.009	−0.023
	(0.004)	(0.013)	(0.034)
English credits gained	0.025	0.117	0.100
	(0.008)	(0.042)	(0.091)
History credits gained	0.007	0.026	0.055
	(0.007)	(0.019)	(0.032)
Total credits gained	0.003	0.002	−0.011
	(0.002)	(0.005)	(0.011)
Average score	0.001	0.001	0.000
	(0.000)	(0.001)	(0.002)
Observations (number of schools)	508	98	36

Notes: The coefficients presented in the table are based on *separate* regressions of the 1999 measurement error on student characteristics, lagged Bagrut outcomes, and school characteristics. The data used are school sample means. Conventional standard errors are presented in parentheses.

characteristics, and therefore the likelihood that the measurement error is correlated with other unobserved confounders is also very low. Admittedly, however, the assumption that the measurement error is not correlated through some unobservables with the change in outcomes from 2000 to 2001 cannot be tested.

Identification based on the random measurement error can be presented formally as follows. Let $S = S^* + \varepsilon$ be the error-affected 1999 matriculation rate used for the assignment, where S^* represents the correct 1999 matriculation rate and ε the measurement error. T denotes participation status, with $T = 1$ for participants and $T = 0$ for nonparticipants. Since $T(S) = T(S^* + \varepsilon)$, once we control for S^*, assignment to treatment is random ("random assignment" to treatment, conditional on the true value of the matriculation rate).

The presence of measurement error creates a natural experiment, where treatment is assigned randomly, conditional on S^*, in a subsample of the 98 eligible schools. Eighteen of the eligible schools had a correct 1999 matriculation rate above the threshold line. Thus, these schools were "erroneously" chosen for the program. For each of these schools, there may have been a school

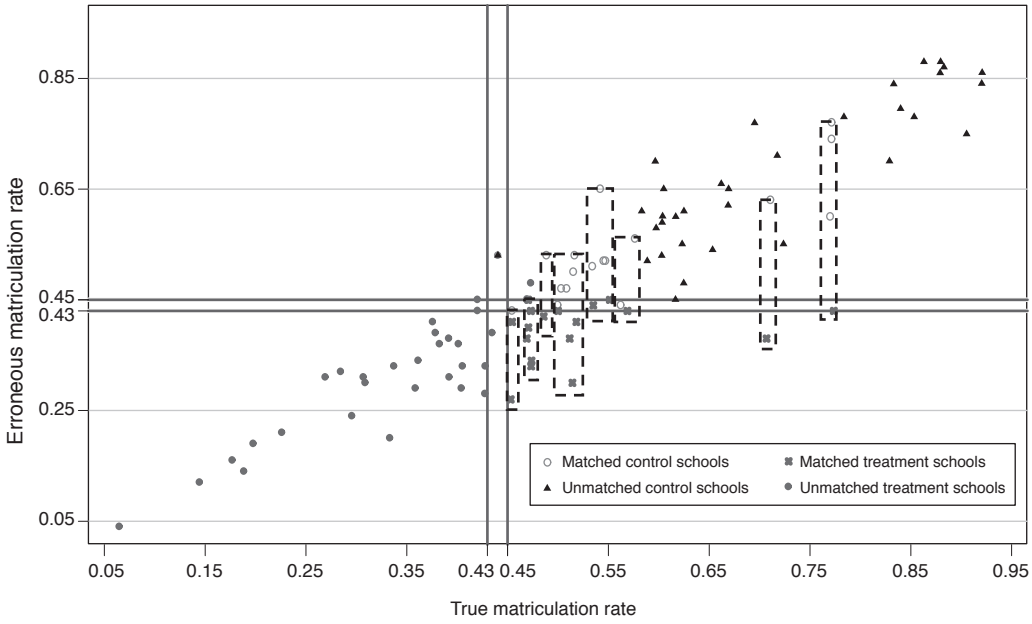

FIGURE 3. DETERMINING THE SAMPLE OF SCHOOLS RANDOMLY ASSSIGNED TO TREATMENT OR CONTROL
(*Sample = 97 schools*)

with an identical correct matriculation rate but with a draw from the (random) measurement error distribution not large (and negative) enough to drop it below the assignment threshold. Such pairing of schools amounts to nonparametrically matching schools on the basis of the value of S^*. I therefore adopted the following matching algorithm: include as a control school all untreated schools that are within $+/-$ 1 percent of the true matriculation rate of any of the 18 erroneously chosen schools. Figure 3 shows the result of this matching procedure where 18 control schools are clustered within rectangles with their treated counterparts. Figure A3 in the Web Appendix presents the more precise matching by connecting the treated-control school pairs within the rectangles. Within each such pairing, assignment to treatment can be viewed as random. Therefore, the 18 untreated schools may be used as a control group that reflects the counterfactual for identification of the effect of the program. Since some of the control schools are matched to more than one treated school, and vice versa, weighted regressions are used to account for the treatment-control differences in sample size within the matched groups (the weights are presented in Table A4 in the Web Appendix). From the point of view of external validity of the results to be presented, it is important to note that the (preprogram) matriculation rates of the 18 treated schools in the RT sample in 2000 span a wide range, from 32 to 79 percent, which overlaps almost completely with the respective range of all the 48 treated schools.

It is important to replicate the regression analysis of the association between the measurement error and school and student characteristics based on this sample of 36 schools. These estimates are reported in column 3 of Table 2, and they are very similar to those based on the sample of all schools and the sample of eligible schools. Two of the 12 estimates are significant but none of the students lagged outcomes in this sample shows any treatment-control imbalance. Actually some of the lagged outcome differences are positive and some negative, which also suggests that any such differences are clearly random.

Table 3—Treatment-Control Balancing Tests: The Randomized Treatment Sample

	2000			2001		
	Treatment (1)	Control (2)	Difference (3)	Treatment (4)	Control (5)	Difference (6)
Panel A. School characteristics						
Religious school	0.330	0.219	0.110	0.324	0.214	0.110
			(0.163)			(0.164)
Arab school	0.158	0.000	0.158	0.155	0.000	0.155
			(0.088)			(0.087)
Lagged Bagrut rate	0.467	0.509	−0.042	0.474	0.475	−0.001
			(0.032)			(0.053)
Two-years lagged Bagrut rate	0.490	0.519	−0.029	0.527	0.528	−0.002
			(0.049)			(0.034)
Panel B. Student background						
Father education	10.685	10.586	0.100	10.539	10.332	0.207
			(0.821)			(0.838)
Mother education	10.624	10.764	−0.140	10.519	10.539	−0.020
			(0.849)			(0.947)
Number of siblings	3.009	2.026	0.983	2.912	1.662	1.250
			(0.410)			(0.384)
Gender (male = 1)	0.513	0.414	0.098	0.556	0.431	0.125
			(0.066)			(0.061)
Immigrant	0.016	0.029	−0.013	0.025	0.012	0.013
			(0.017)			(0.018)
Asia-Africa ethnicity	0.218	0.325	−0.107	0.235	0.276	−0.041
			(0.062)			(0.054)
Panel C. Student lagged outcomes						
Math credits gained	0.337	0.277	0.061	0.256	0.453	−0.197
			(0.172)			(0.118)
English credits gained	0.155	0.077	0.078	0.107	0.079	0.028
			(0.051)			(0.061)
Total credits attempted	5.251	4.594	0.657	5.322	5.342	−0.020
			(0.674)			(0.498)
Total credits gained	4.308	3.761	0.547	4.218	4.482	−0.264
			(0.601)			(0.393)
Average score	63.131	64.774	−1.643	62.121	67.710	−5.589
			(2.591)			(2.217)
Observations	2,654	2,369	5,023	2,598	2,236	4,834
Observations, weighted	4,095	3,818	7,913	3,812	3,679	7,491
Number of schools	18	18	36	18	18	36

Notes: Standard errors in parentheses are adjusted for school-level clustering. Observations were weighted with frequency weights in order to have similar number of students in control and treatment schools within each group of schools with close true matriculation rate. The schools status of nationality and religiosity does not change. Any change in the means across years reflects relative changes in the number of students in a cohort. This table is based on the math sample.

Table 3 presents the preprogram (2000) and postprogram (2001) means of school and student characteristics for the 18 treated schools and the 18 control schools in the RT sample. The first panel compares the school-level covariates. The treatment-control differences and their standard errors are presented in columns 3 and 6. Treatment and control groups are balanced in terms of religious status but not in terms of nationality, since there are no Arab schools in the control

group. The one- and two-year lagged matriculation rates are perfectly balanced for the 2001 seniors cohort, as expected, since we used the 1999 S^* to match control to treated schools in this sample. The second panel presents the balancing tests for student characteristics. There are differences in the number of siblings and in the gender composition, but no significant differences in parental schooling, immigrant status, and ethnicity. For example, the mean of mother's schooling in 2001 is identical in the two groups, a remarkable contrast to the respective difference of 2.3 years schooling estimated based on the eligible sample (Table 1). The third panel presents balancing tests for students' lagged outcomes, accumulated credits in Bagrut exams during 10th and 11th grade, which for the 2001 cohort should be viewed as *preprogram outcomes*. Most of these differences are small and none is significant except for the lagged average score in 2001. Of particular interest are the lagged credits in math and English. No significant treatment-control differences are observed for these outcomes in either year. The means of total lagged credits, which includes credits in all subjects for which students were tested during 10th and 11th grades, is also identical for the two groups in both years.

The evidence presented in Table 3 demonstrates balance between treatment-control schools in most student and school characteristics, in sharp contrast to the respectively large differences seen in Table 1. The almost perfect equality of the two groups in terms of lagged outcomes (presented in the lower panel of Table 3), strengthens even further the case for the comparison group produced by the natural experiment. As will be shown in the next section, there is also almost perfect equality in the 12th grade outcomes of the cohort that graduated before treatment: the testing rate and the pass rate in math and English are practically identical for the preprogram graduating cohort in the control and treatment schools. This evidence is also reassuring in terms of the causal interpretation of the posttreatment differences in these outcomes, which is investigated in the next section.

B. *Estimation and Results*

The following model was used as the basis for regression estimates using the RT sample:

$$Y_{ijt} = \alpha + \mathbf{X}'_{ijt}\beta + \mathbf{Z}'_{jt}\gamma + \delta T_{jt} + \Phi_j + \eta D_t + \varepsilon_{ijt}, \tag{1}$$

where i indexes students, j indexes schools, t indexes years 2000 and 2001, T is the assigned treatment status, $\mathbf{X}$ and $\mathbf{Z}$ are vectors of student and school level covariates, and D_t denotes year effects with a factor loading η. The treatment indicator T_{jt} is equal to the interaction between a dummy for treated schools and a dummy for the year 2001. The regressions were estimated using pooled data from both years (the two adjacent cohorts of 2000 and 2001), stacked as school panel data with fixed school-level effects (Φ_j) included in the regression. The resulting estimates can be interpreted as a student-weighted difference-in-differences procedure comparing treatment effects across years. The estimates are implicitly weighted by the number of students in each school. The introduction of school fixed effects controls for time-invariant omitted variables and also provides an alternative control for school-level clustering. However, standard errors are also clustered at the school level for each cross section. The school fixed effects also absorb some of the variability in average matriculation outcome rates by school, possibly leading to a gain in precision.

Table 4 presents the results from estimating equation (1) with the RT sample. I present results for two specifications, the first with a limited set of controls which includes only school fixed effects, year dummies, the one- and two-year lagged matriculation rate, and the number of attempted credit units. The latter variable accounts for idiosyncratic variation in the matriculation testing program across schools from year to year. The second specification includes in addition the following individual characteristics as controls: number of siblings, gender dummy,

TABLE 4—DID ESTIMATES OF THE EFFECT OF TEACHER BONUSES ON MATH AND ENGLISH OUTCOMES BASED ON THE RANDOMIZED TREATMENT SAMPLE

	Math					
	All quartiles		Estimates by quartile			
	Limited control (1)	Full control (2)	1st (3)	2nd (4)	3rd (5)	4th (6)
Testing rate						
Control group mean	0.802		0.419	0.815	0.903	0.971
Treatment effect	0.046	0.041	0.133	0.055	0.037	−0.021
	(0.027)	(0.021)	(0.051)	(0.035)	(0.021)	(0.029)
	[0.038]	[0.029]	[0.068]	[0.047]	[0.030]	[0.039]
Pass rate						
Control group mean	0.637		0.258	0.503	0.726	0.928
Treatment effect	0.110	0.087	0.146	0.209	0.106	−0.026
	(0.036)	(0.028)	(0.048)	(0.063)	(0.035)	(0.029)
	[0.051]	[0.040]	[0.065]	[0.087]	[0.047]	[0.041]
Average score						
Control group mean	55.046		21.232	46.917	63.946	77.710
Treatment effect	5.469	5.307	9.798	10.920	6.352	−0.861
	(2.292)	(1.950)	(3.497)	(4.104)	(2.122)	(2.493)
	[3.249]	[2.739]	[4.768]	[5.686]	[2.927]	[3.443]
Conditional treatment effect						
Passing rate		0.052	0.051	0.161	0.073	−0.007
Proportion of unconditional effect		59%	35%	77%	69%	–
Average score		2.323	2.465	7.006	3.541	0.839
Proportion of unconditional effect		44%	25%	64%	56%	–
Observations	9,857		2,421	2,365	2,424	2,647
	English					
	All quartiles		Estimates by quartile			
	Limited control (7)	Full control (8)	1st (9)	2nd (10)	3rd (11)	4th (12)
Testing rate						
Control group mean	0.865		0.529	0.903	0.972	0.977
Treatment effect	0.040	0.033	0.129	0.013	0.004	−0.003
	(0.017)	(0.013)	(0.045)	(0.024)	(0.014)	(0.019)
	[0.025]	[0.019]	[0.060]	[0.033]	[0.019]	[0.026]
Pass rate						
Control group mean	0.795		0.455	0.770	0.906	0.959
Treatment effect	0.047	0.039	0.107	0.071	−0.011	−0.009
	(0.022)	(0.020)	(0.040)	(0.035)	(0.021)	(0.034)
	[0.031]	[0.028]	[0.052]	[0.048]	[0.028]	[0.045]
Average score						
Control group mean	59.496		35.464	59.608	68.901	73.751
Treatment effect	3.240	2.527	5.889	1.790	−0.648	0.344
	(1.666)	(1.452)	(2.295)	(2.840)	(1.708)	(2.077)
	[2.359]	[2.040]	[3.150]	[3.932]	[2.341]	[2.828]
Conditional treatment effect						
Passing rate		0.009	−0.001	0.059	−0.015	−0.006
Proportion of unconditional effect		23%	–	84%	–	–
Average score		0.238	−2.255	0.901	−0.915	0.550
Proportion of unconditional effect		9%	–	50%	–	–
Observations	10,111		2,506	2,390	2,500	2,715

Notes: Standard errors in parentheses are clustered at the school-year-combination level. Standard errors in brackets are clustered at the school level. Observations were weighted with frequency weights in order to have similar number of students in control and treatment schools within each group of schools with close true matriculation rate. In columns 3–6 and 9–12, treatment effects vary by quartiles of previous test score distribution. The estimates are taken from four separate regressions, one for each of the quartiles. Student level controls—in all columns except 1 and 7—include a set of dummy variables for the number of siblings and father's and mother's education, the school's one-year and two-years lagged mean matriculation rate, a dummy for Asia-Africa ethnic background, immigration status, gender dummy, the number of credit units *attempted*, the average score in those attempted units, overall credit units *awarded*, and credit units awarded for the subject in question only. School fixed effects are included in all specifications. In columns 1 and 7, the controls are school dummies, school's one-year and two-years lagged matriculation rate, and student's attempted credit units. Control group mean row shows the mean for students in control schools in 2001.

father's and mother's education, a dummy indicator for immigration status, a dummy variable indicating Asian/African ethnicity, number of credits gained in the relevant subject (i.e., math or English) before treatment, total credits accumulated before treatment, and their average score.

The estimated treatment effect based on the limited control specification in math (column 1) and English (column 7) is positive and significant for all three outcomes. Estimates based on the full specification are presented in columns 2 and 8. These estimates are very close to those of the limited specification, confirming what is expected given the balancing tests presented in Table 3. I therefore discuss only the point estimates of the full specification.[14] The effect of treatment on test taking in math is 0.041, a 5 percent improvement relative to the mean of the control schools (0.802). The effect of treatment on the pass rate is 0.087, a relative gain of 14 percent. The treatment effect on the math average score represents a 10 percent improvement. In English the relative gains due to treatment are more modest: 4 percent improvement in the testing rate, 5 percent in the pass rate, and 4 percent for the average score.

The interpretation of these gains as causal is based on the random assignment of program status by measurement error, conditional on the actual 1999 matriculation rate. This conditioning was achieved by matching treated schools to control schools with exactly the same 1999 matriculation rate (S^*) in the RT sample. As a result of this matching, the high negative correlation (-0.689) between S^* and T estimated in the eligible sample was reduced to practically zero among these variables in the RT sample. So conditioning nonparametrically on S^* makes treatment random in the RT sample. Nevertheless, I included in the stacked data model the 1999 S^* as a control for the 2001 cohort outcomes and, given the symmetry of the stacked panel data structure, I also use the 1998 S^* as a control for the 2000 cohort. Omitting this control lowers the treatment-effect estimates for the testing rate for math to 0.037 (from 0.041) and for English to 0.029 (from 0.033) and increases the estimated effect on the math pass rate to 0.092 (from 0.087), but leaves the other estimated effects basically unchanged. Thus, including the two-year lagged true matriculation rate in the equation does not change the treatment effect estimates much because it is not correlated with treatment status in the RT sample, and because its coefficient in the outcome equation is not significant (see Web Appendix Table A1). Dropping the one-year lagged matriculation rate from the equation also left the estimates unchanged.

The analysis thus far was based on the RT sample with panel data that allowed for difference-in-differences estimation. In Table 5, I report cross-sectional estimates for the 2000 and 2001 RT samples in order to examine how sensitive the treatment effect estimates are to controlling for school fixed effects. Column 1 and 6 present the estimates based on the full sample and full control specification for math and English, respectively. Panel A presents the results for 2001 and panel B presents those for 2000. These estimates show that most of the estimated gain reflected in the stack estimates of Table 4 (presented in columns 2 and 8) is due to positive and significant effects in the estimates of 2001 since the estimates for 2000 are mostly small, and sometimes even negative. For example, the effect on the math testing rate in 2001 is 0.030 and in 2000 it is -0.022, and the difference between the two is 0.052, similar to the 0.041 respective estimate reported in column 2 of Table 4. However, in most cases the difference-in-differences estimates presented in Table 4 are marginally larger and much more precise than the cross-sectional estimates presented in Table 5, suggesting that the school fixed effects are necessary in order to

[14] Table A1 in the Web Appendix reports the estimated coefficients of all the variables included in the fully specified math and English equations for all three outcomes in each subject. I do not report the coefficient estimates on father's and mother's years of schooling, however, because these are entered into the equation as a set of dummy variables (a separate indicator for each level of years of schooling). This specification was preferred in order to avoid imputation of parental schooling for missing values. Students with missing father's or mother's schooling are grouped under two separate dummy indicators. The estimates of the treatment effects are unchanged, however, when continuous parental schooling variables replace the parental schooling dummy variable indicators.

TABLE 5—CROSS-SECTION ESTIMATES OF THE EFFECT OF TEACHER BONUSES ON MATH AND ENGLISH OUTCOMES BASED ON THE RANDOMIZED TREATMENT SAMPLE

	Math					English				
		Estimates by quartile					Estimates by quartile			
	All quartiles (1)	1st (2)	2nd (3)	3rd (4)	4th (5)	All quartiles (6)	1st (7)	2nd (8)	3rd (9)	4th (10)
2001										
Testing rate										
Control group mean	0.802	0.419	0.815	0.903	0.971	0.865	0.529	0.903	0.972	0.977
Treatment effect	0.030	0.092	0.087	0.011	−0.096	0.026	0.076	0.054	−0.015	−0.043
	(0.032)	(0.051)	(0.028)	(0.025)	(0.037)	(0.020)	(0.037)	(0.019)	(0.016)	(0.025)
Pass rate										
Control group mean	0.637	0.258	0.503	0.726	0.928	0.795	0.455	0.770	0.906	0.959
Treatment effect	0.037	0.089	0.139	−0.005	−0.127	0.025	0.063	0.068	−0.016	−0.057
	(0.052)	(0.048)	(0.070)	(0.062)	(0.045)	(0.028)	(0.047)	(0.033)	(0.024)	(0.029)
Average score										
Control group mean	55.046	21.232	46.917	63.946	77.710	59.496	35.464	59.608	68.901	73.751
Treatment effect	2.561	4.818	10.216	−0.338	−9.327	1.769	3.196	4.328	−1.019	−2.953
	(3.706)	(3.489)	(4.745)	(4.155)	(4.154)	(1.947)	(2.894)	(2.053)	(2.054)	(2.666)
Observations	4,834	1,234	1,161	1,196	1,243	4,964	1,281	1,163	1,227	1,293
2000										
Testing rate										
Control group mean	0.786	0.492	0.771	0.909	0.945	0.846	0.587	0.834	0.974	0.964
Treatment effect	−0.022	−0.027	0.055	−0.018	−0.057	−0.005	0.015	0.066	−0.019	−0.037
	(0.022)	(0.055)	(0.040)	(0.023)	(0.021)	(0.019)	(0.045)	(0.042)	(0.015)	(0.016)
Pass rate										
Control group mean	0.654	0.305	0.619	0.785	0.880	0.744	0.438	0.717	0.870	0.926
Treatment effect	−0.023	0.008	−0.005	−0.034	−0.072	0.002	0.000	0.067	0.032	−0.060
	(0.035)	(0.047)	(0.046)	(0.037)	(0.036)	(0.038)	(0.073)	(0.052)	(0.031)	(0.020)
Average score										
Control group mean	51.895	24.490	46.606	62.379	72.783	56.821	34.430	53.404	66.227	72.019
Treatment effect	0.185	−0.274	4.900	0.866	−4.372	1.929	1.828	8.422	2.505	−2.375
	(2.537)	(3.293)	(3.159)	(2.891)	(3.267)	(2.519)	(4.615)	(4.018)	(1.898)	(1.715)
Observations	5,023	1,187	1,204	1,228	1,404	5,147	1,225	1,227	1,273	1,422

Notes: Standard errors in parentheses are clustered at the school level. Observations were weighted with frequency weights in order to have a similar number of students in control and treatment schools within each group of schools with close true matriculation rate. In columns 2–5 and 7–10, treatment effects vary by quartiles of previous tests score distribution. The estimates are taken from four separate regressions, one for each of the quartiles. Student-level controls include a set of dummy variables for the number of siblings and father and mother education, the school's lagged mean matriculation rate, a dummy for Asia-Africa ethnic background, immigration status, gender dummy, the number of credit units attempted, the average score in those *attempted* units, overall credit units *awarded*, and credit units awarded for the subject in question only. School fixed effects are included in all specifications.

produce consistent and precise estimates. However, as will be shown below, this is much less the case when the treatment effect estimates are allowed to vary by student ability.

C. *Allowing for Heterogeneity in the Effect of Treatment by Student Ability*

As an additional check of the causal interpretation of the results presented in columns 2 and 7 of Table 4, I estimated models that allow treatment effects to vary with lagged outcomes. In particular, I allowed for an interaction of the treatment effect with the mean credit-weighted average score on all previous matriculation exams (coding zeros for those who had taken no exams). Using this average score, which is a powerful predictor of students' success in the math

and English tests, I coded dummies for each quartile of the score distribution. Using the quartile dummies, it is possible to estimate the following model for each of the three outcomes of interest in English and math:

$$Y_{ijt} = \alpha + \mathbf{X}'_{ijt}\beta + \mathbf{Z}'_{jt}\gamma + \Sigma_q d_{qi}\mu_q + \Sigma_q \delta_q d_{qi} T_{jt} + \Phi_j + \eta D_t + \varepsilon_{ijt}, \quad (2)$$

where d_{qi} $(q = 2, 3, 4)$ indicates the quartile of a student's credit-unit-weighted average test score on tests taken before January 2001, when the program was implemented, δ_q is a quartile-specific treatment effect, and μ_q is a quartile main effect. An alternative strategy is to permit the effect of all variables to differ by quartile, and this is the specification I preferred, although the results are not very different when equation (2) is estimated instead.

Students with very high lagged scores were likely to be able to take and pass the exams in each of the subjects without the help of the program. This claim is supported by the fact that the mean matriculation rate in this quartile was 90 percent in 2000. The teachers may have realized by themselves that the margin for improvement for top quartile students is very limited, and therefore they may have directed their efforts toward the lower quartiles that had larger scope for improvement. Therefore, one would not expect to find an effect of the teacher incentive program on students in the top quartile. In contrast, students with scores around or below the mean of the score distribution fell into a range in which extra effort—by their teachers and by themselves—may have made a difference. Therefore, I looked for significant estimates for students mainly in two lowest quartiles.

In columns 3–6 and 9–12 of Table 4, I report the results of estimating the treatment effect of the teacher incentive program for math and English, respectively, based on stratified samples by quartiles. The pattern in this table suggests that the average effects on math outcomes reported in column 2 of Table 4 originate from the effects in the first three quartiles—students just above and below the average—while no significant effects were estimated for the top students (quartile 4). The average effect on English outcomes originates from the effect in the first two quartiles only. The zero effect in the fourth quartile is not surprising, given that most students in this quartile were expected to take the exams (as evidenced by the control group mean in this quartile, 97 percent in math and 98 percent in English) as scheduled and pass (93 percent in math and 96 percent in English). Another interesting pattern to note in these columns is that the estimated effect on the testing rate is large and significant only for the first quartile, while the effect in the second and third quartiles is mainly on the pass rate and on the average test score, with no effect on the test taking rate. In the third quartile for English there is no significant increase in any of the outcomes, probably due to the very high mean outcomes of this group, which interestingly are much closer to the means of the fourth quartile in this subject than the respective third quartile in math outcomes.

The improvement in all three outcomes in the first quartile is quite large relative to the control group mean: the gain in the testing rate implies a 32 percent increase in math and 24 percent in English, the pass rate is up by 57 percent in math and 23 percent in English, while the increase in the mean test score is 46 percent in math and 17 percent in English. The gains in math are much larger than in English, but the improvements in the latter are still sizeable. A similar pattern is exhibited in the second quartile. For example, we see a 42 percent increase in the pass rate and a 23 percent increase in the average score in math, versus a 9 and 3 percent increase implied by the respective estimates in the second quartile in English.

It is important to note that the evidence that most of the effect of the program is concentrated among students below the median of the ability distribution can be related to the specifics of the incentive scheme of this program. Particularly important is the feature that two outcome measures were used to rank teachers, namely the pass rate and the average score, and that a larger

weight (25 percent higher) was assigned to the pass rate in computing each teacher's composite rank.[15] Not only is the effect concentrated below the median, but the relative effect on the pass rate is much larger than the effect on the test score. In math, for example, the effects on the pass rate in the first and second quartiles constitute a 57 and 42 percent improvement, respectively, while the gain in the average test scores in these two quartiles was only 46 and 23 percent. The same pattern of relative gain is seen in English, although the differences across the two outcomes are not as large. The fact that the improvement in outcomes was mainly below the median and that it was much larger in the pass rate suggests that teachers understood the monetary advantage of improving the pass rate relative to improvement in the average test score, and that they were aware that there is much more potential for improvement in both outcomes among pupils below the median of the students' ability distribution.[16]

As discussed earlier with respect to the results regarding the average treatment effects, I also estimated cross-sectional estimates for the 2000 and 2001 quartile samples, and these are presented in columns 2–5 (for math) and columns 7–10 (for English) of Table 5. These estimates show much more clearly than the full sample results (presented in columns 1 and 6 of Table 5) that most of the estimated gains reflected in the stack estimates of Table 4 are due to positive and significant effects in the estimates of the 2001 cross section, since the 2000 cross-section estimates are mostly small, sometimes even negative and not significant. Focusing on the estimates for the second quartile, the estimated effect on the math pass rate in 2000 is -0.005 (s.e. 0.046) while in 2001 it is 0.139 (s.e. 0.070), and the respective stack estimate in Table 4 is 0.209, not far from the simple difference of the 2001 and 2000 cross-section estimates. This pattern is seen for the other outcomes as well, although not all of the cross-sectional estimates align so perfectly in comparison to their respective stack DID estimates that are presented in Table 4.

D. *Estimating the Effects on the Conditional Pass Rate and Test Score*

The results presented above show that changes in teacher incentives increased the exam taking rate and the unconditional pass rate and test score. The latter may reflect an impact through two channels. The first is through a change in the exam taking rate which may increase if students, who otherwise would not take the exam, take it due to the program. This channel will increase the unconditional pass rate "mechanically" as long as some of these students obtain a passing score. The second channel is through an increase in the conditional pass rate among students who would have taken the exam regardless of the program. A back-of-the-envelope calculation of the treatment effect on the conditional pass rate and average test score can be derived as follows: let Y_i denote the outcome of school i and let P_i be the fraction of students in school i who take the exam. School's i expected outcome is equal to

$$E(Y_i) = P_i Y_i^1 + (1 - P_i)\, Y_i^0 = P_i\, Y_i^1,$$

where Y_i^1 is the average pass rate among students in school i who took the exam, and Y_i^0 is the respective average of students in the same school who did not take the exam. Given the tournament scheme (see Section I) which assigns the value 0 for all students who choose not to take the exam, school i's average expected pass rate is equal to $E(Y_i) = P_i Y_i^1$. The average overall

[15] This differential weighting scheme was computed by awarding a 25 percent increase in points for each ranking (first to fourth place) based on the pass rate. More details are given in Section IIB above.

[16] A recent paper (Derek Neal and Diane Whitmore Schanzenbach 2007) shows a similar result: teachers responding to the accountability system in Chicago focused time and resources on students just on the margin of passing the relevant exam.

treatment effect (*ATE*), assuming that schools were randomly assigned into treatment and control group, is equal to

$$ATE = P_1 Y_1^1 - P_0 Y_0^0 = (P_1 - P_0) Y_1^1 + (Y_1^1 - Y_0^1) P_0,$$

where P_1 and P_0 are the test taking rates and Y_1^1 and Y_0^1 are the conditional pass rates at the treated and control schools, respectively. The *ATE* could also be expressed in terms of the unconditional outcomes, $ATE = Y_1 - Y_0$, where Y_1 is the unconditional test pass rate, that is, $Y_1 = P_1 Y_1^1$ and $Y_0 = P_0 Y_0^1$.

The overall treatment effect can therefore be decomposed into the average treatment effect due to the increase in the test taking rate, $(P_1 - P_0) Y_0^1$, and the average treatment effect conditional on the test taking rate, $(Y_1^1 - Y_0^1)P_0$. The lower panel in Table 4 presents the results of this decomposition for the pass rate and for the average test score in English and math, based on the estimates presented in the top panel of the table. Of the 8.7 percent increase in the unconditional math pass rate, 5.2 percent is due to an increase in the conditional pass rate, therefore accounting for 59 percent of the unconditional improvement in the pass rate. The other 3.5 percent improvement resulted from an increase in the test participation rate. A similar analysis based on the estimates by quartile leads to similar results, although the proportion of the unconditional effect is higher in the second (77 percent) and third (69 percent) quartiles because there is almost no effect on the testing rate, and lowest in the first quartile (35 percent). Similar decomposition of the effect on the average score implies that on average almost half of the increase in the unconditional test score gain is due to an increase in the conditional test score. The pattern across quartiles is similar to that of the pass rate, being highest in the second and third quartiles (64 percent and 56 percent, respectively) and lowest in the first (25 percent).

The decomposition of the unconditional English treatment effects suggests that the increase in the exam taking rate was more important in this subject, accounting for 23 percent of the overall gain in the pass rate. However, in the second quartile the conditional pass rate contributed 84 percent to the overall gain. The gain in the English test score is due primarily to the increase in the testing rate in this subject, but in the second quartile the conditional increase still accounted for half of the overall test score gain. However, it should again be noted that these estimates cannot be viewed as causal because test taking is an endogenous outcome and therefore estimates of the effects on the conditional pass rate and average score may reflect selection bias. But this decomposition method in all likelihood yields a lower bound to the contribution of the conditional outcomes to the overall change, as students who were induced to take the test by the program are likely to be weaker than students who would have taken the test regardless of the program.[17]

E. *Alternative Identification Methods*

To check the robustness of the results based on the RT sample, I also used two additional alternatives methods to identify the effect of the teacher bonus program. The first is an RD design. Given that the rule governing selection to the program was simply based on a discontinuous function of a school observable, the probability of receiving treatment changes discontinuously as a function of this observable. The discontinuity in our case is a sharp decrease (to zero) in the probability of treatment beyond a 45 percent school matriculation rate for nonreli-

[17] The decomposition of the unconditional treatment effect can also be carried out in the opposite way, namely, by using the following equation: $ATE = P_1 Y_1^1 - P_0 Y_0^0 = (P_1 - P_0) Y_0^1 + (Y_1^1 - Y_0^1) P_1$. However, this equation provides an upper bound for the conditional estimates because $(Y_1^1 - Y_0^1) P_0 < (Y_1^1 - Y_0^1) P_1$.

gious Jewish schools and beyond 43 percent for Jewish religious schools and Arab schools.[18] I exploited this sharp discontinuity to define a treatment sample that included schools that were just below (up to -5 percent) the threshold of selection to the program and a comparison group that included untreated schools that were just above (up to $+5$ percent) this threshold. The time series on school matriculation rates show that the rates fluctuate from year to year for reasons that transcend trends or changes in the composition of the student body. Some of these fluctuations are random. Therefore, marginal (in terms of distance from the threshold) participants may be similar to marginal nonparticipants. The degree of similarity depends on the width of the band around the threshold. Sample size considerations exclude the possibility of a bandwidth lower than 10 percent, and a wider band implies fluctuations of a magnitude that are not likely to be related to random changes. Therefore, a bandwidth of about 10 percent seems to be a reasonable choice in our case. The main drawback of this approach is that it produces an estimate from marginally (relative to the threshold) exposed schools only. However, this sample may be of particular interest because the threshold schools could be representative of the schools that such programs are most likely to target.

There are 13 untreated schools with matriculation rates in the 0.46–0.52 range and 14 treated schools in the 0.40–0.45 range (Figure A4 in the Web Appendix).[19] The 0.40–0.52 range may be too large, but I can control for the value of the assignment variable (the mean matriculation rate) in the analysis. Note, also, that there is some overlap between this sample and the natural experiment sample. Eleven of the 14 treated schools and 8 of the 11 control schools in the RD sample are also part of the RT natural experiment sample, leaving only six schools (3 control and 3 treated), which are included in the former but not in the latter. However, there are 17 schools in the RT sample (7 treated and 10 control) that are not included in the RD sample, which suggests that there is enough "informational value added" in each of the samples.

Table 6 replicates Table 3 for the RD sample. The treatment-control differences and standard errors in the student background variables (columns 3 and 6) reveal that the two groups are very similar in both years in all characteristics, except the ethnicity variable in year 2000 and number of siblings in 2001. However, both estimated differences are only marginally significant. The third panel reveals some treatment-control differences: in math lagged credits and in the average score for the 2001 cohort, and in English lagged credits for the 2000 cohort. But the control-treatment gaps in lagged credits are opposite in sign in math (negative) and English (positive) and in each subject they are significant for only one of the two cohorts.

Table 7 presents the results from the estimation of equations (1) and (2) using the RD sample.[20] The treatment-effect estimates are all positive and significantly different from zero for all the sample mean English and math outcomes. The estimates with the limited set of controls are very similar to those with all the controls included. Focusing on the estimates based on the fully specified equations (columns 2 and 8), it can be seen that they are qualitatively similar to those presented in Table 4 in the respective columns, with one important difference. Both in math and English the RD estimates of the effect on the average testing rate are higher than those obtained based on the RT sample. In English, for example, the effect on the testing rate in the RD sample

[18] The Ministry of Education decided on a lower threshold for Jewish religious schools and Arab schools in order to lower their proportion among the treated schools.

[19] The figure shows that there is one treated school with an erroneous 1999 matriculation rate of 48 percent, and therefore it should have not been included in the program. I do not include this school in the RD sample because it is out of the range of the RD treated sample (40–45 percent).

[20] In principle, the identification based on the RD method is conditioned on controlling for the erroneously measured matriculation rate that was actually used to assign schools to the program. However, the fixed school-level effects, which are included in each regression, control for the 1999 erroneously measured matriculation rate and therefore the latter is dropped from the equation.

TABLE 6—TREATMENT-CONTROL BALANCING TESTS: THE REGRESSION DISCONTINUITY SAMPLE

	2000			2001		
	Treatment (1)	Control (2)	Difference (3)	Treatment (4)	Control (5)	Difference (6)
Panel A. School characteristics						
Religious school	0.100	0.301	−0.201 (0.142)	0.095	0.290	−0.195 (0.140)
Arab school	0.131	0.000	0.131 (0.094)	0.132	0.000	0.132 (0.096)
Lagged "Bagrut" rate	0.448	0.495	−0.047 (0.017)	0.458	0.470	−0.012 (0.041)
Panel B. Student background						
Father education	11.027	10.219	0.808 (0.591)	10.835	10.081	0.753 (0.643)
Mother education	11.095	10.526	0.570 (0.659)	11.027	10.527	0.501 (0.711)
Number of siblings	2.622	2.288	0.335 (0.352)	2.605	1.902	0.703 (0.383)
Gender (male = 1)	0.493	0.425	0.068 (0.058)	0.499	0.451	0.048 (0.052)
Immigrant	0.014	0.045	−0.031 (0.021)	0.013	0.009	0.004 (0.007)
Asia-Africa ethnicity	0.215	0.313	−0.097 (0.052)	0.214	0.273	−0.060 (0.054)
Panel C. Student lagged outcomes						
Math credits gained	0.185	0.364	−0.180 (0.131)	0.185	0.452	−0.267 (0.128)
English credits gained	0.207	0.053	0.155 (0.061)	0.183	0.101	0.083 (0.088)
Total credits attempted	4.788	4.944	−0.156 (0.476)	5.064	5.346	−0.283 (0.489)
Total credits gained	4.008	4.066	−0.058 (0.376)	4.188	4.394	−0.206 (0.384)
Average score	61.671	64.548	−2.877 (2.932)	61.797	65.770	−3.973 (1.973)
Observations	2,471	1,638	4,109	2,401	1,519	3,920
Number of schools	14	13	27	14	13	27

Notes: Standard errors in parentheses are adjusted for school-level clustering. The schools status of nationality and religiosity does not change. Any change in the means across years reflects relative changes in the number of students in a cohort. This table is based on the math sample.

is almost 50 percent higher than the respective RT estimate. This difference is largely due to the positive estimated effects on test taking in the third and fourth quartiles, which were totally absent in the RT sample results. On the other hand, the estimated effects on the pass rate and the test score in the first and second quartiles based on the two methods are very similar.

To interpret the differences between the estimates presented in Tables 4 and 7, it is important to note that the main conceptual difference between the RT and the RD methods is that the latter does not control for S^* and, if there were no measurement errors, the RD design would have compared, in addition, pupils or schools with different S^*. The two methods would therefore yield similar results if either S^* is weakly related to the outcomes or if the variance of the

TABLE 7—DID ESTIMATES OF THE EFFECT OF TEACHER BONUSES ON MATH AND ENGLISH OUTCOMES BASED ON THE REGRESSION DISCONTINUITY SAMPLE

	Math					
	All quartiles		Estimates by quartile			
	Limited control (1)	Full control (2)	1st (3)	2nd (4)	3rd (5)	4th (6)
Testing rate						
Control group mean	0.767		0.407	0.775	0.880	0.947
Treatment effect	0.072	0.055	0.112	0.027	0.034	0.031
	(0.034)	(0.029)	(0.068)	(0.051)	(0.037)	(0.024)
	[0.049]	[0.040]	[0.090]	[0.070]	[0.052]	[0.032]
Pass rate						
Control group mean	0.602		0.248	0.503	0.699	0.889
Treatment effect	0.111	0.088	0.091	0.141	0.086	0.056
	(0.037)	(0.028)	(0.049)	(0.056)	(0.035)	(0.026)
	[0.052]	[0.040]	[0.064]	[0.077]	[0.049]	[0.036]
Average score						
Control group mean	51.219		20.388	44.620	60.009	75.259
Treatment effect	6.733	5.790	4.408	4.687	6.608	6.293
	(2.415)	(1.812)	(2.710)	(3.299)	(2.054)	(2.566)
	[3.437]	[2.555]	[3.711]	[4.636]	[2.845]	[3.536]
Conditional treatment effect						
Passing rate		0.040	0.010	0.118	0.055	0.026
Proportion of unconditional effect		46%	11%	84%	64%	47%
Average score		1.752	−2.037	2.785	3.968	3.743
Proportion of unconditional effect		30%	–	59%	60%	59%
Observations	8,029		2,002	1,983	2,032	2,012

	English					
	All quartiles		Estimates by quartile			
	Limited control (7)	Full control (8)	1st (9)	2nd (10)	3rd (11)	4th (12)
Testing rate						
Control group mean	0.826		0.489	0.859	0.939	0.955
Treatment effect	0.053	0.048	0.092	0.028	0.028	0.040
	(0.022)	(0.018)	(0.066)	(0.025)	(0.013)	(0.017)
	[0.032]	[0.026]	[0.089]	[0.035]	[0.018]	[0.023]
Pass rate						
Control group mean	0.745		0.377	0.725	0.882	0.923
Treatment effect	0.039	0.033	0.109	0.009	−0.041	0.036
	(0.029)	(0.022)	(0.071)	(0.039)	(0.018)	(0.027)
	[0.041]	[0.031]	[0.096]	[0.055]	[0.025]	[0.036]
Average score						
Control group mean	55.243		30.143	56.915	65.432	69.825
Treatment effect	2.975	2.671	4.144	−1.530	0.709	4.419
	(1.858)	(1.421)	(3.600)	(2.762)	(1.446)	(1.826)
	[2.642]	[2.000]	[5.014]	[3.856]	[2.011]	[2.462]
Conditional treatment effect						
Passing rate		−0.011	0.027	−0.017	−0.067	−0.002
Proportion of unconditional effect		–	25%	–	–	–
Average score		−0.712	−1.886	−3.482	−1.295	1.341
Proportion of unconditional effect		–	–	–	–	30%
Observations	8,264		2,065	2,066	2,066	2,067

Notes: Standard errors in parentheses are clustered at the school-year-combination level. Standard errors in brackets are clustered at the school level. In columns 3–6 and 9–12 treatment effects vary by quartiles of previous test score distribution. The estimates are taken from a four separate regressions, one for each of the quartiles. Student level controls—in all columns except 1 and 7—include a set of dummy variables for the number of siblings and father's and mother's education, the school's (one-year) lagged mean matriculation rate, a dummy for Asia-Africa ethnic background, immigration status, gender dummy, the number of credit units *attempted*, the average score in those attempted units, overall credit units *awarded*, and credit units awarded for the subject in question only. School fixed effects are included in each model. In columns 1 and 7, the controls are school fixed effects, school's one-year lagged matriculation rate, and student's attempted credit units. Control group mean row shows the mean for students in control schools in 2001.

measurement error is large relative to the variance of S^* around the cutoff point (which means that those above and below the critical S have approximately the same S^*). The first condition is met, as S^* has very small positive correlations with the three outcomes in each subject. Actually, the highest correlation observed is between S^* and the testing rate in English (0.043) and in math (0.069); indeed, for these outcomes we observed the largest differences between the RT and the RD estimates. The second condition is not met because within the range (0.40-0.52) around the cutoff point of the assignment variable (S), the two relevant variances are very similar, 0.075 for the measurement error and 0.078 for S^*.

The second alternative identification method used is a difference-in-differences estimation strategy (DID) that makes use of the panel data (before and after) of all the eligible schools, and relies on school fixed effects to overcome the treatment-control imbalances in the eligible schools sample in Table 1. The DID treatment effect estimates are presented in Web Appendix Table A2. They are positive and significantly different from zero, but they are much lower than the RT estimates, suggesting that they are significantly downward biased. A discussion of these differences is presented in the Web Appendix.

III. Do Teachers' Pedagogy and Effort Respond to Financial Incentives?

The evidence in the previous section clearly shows that the teacher incentive program led to significant improvements in student achievement in English and math. How closely do these improvements correspond to greater effort on the part of teachers? Do they reflect different pedagogy? The answers to these questions may shed some light on the concern that financial incentives may mainly affect teachers' efforts to prepare students for tests, in what is often termed "teaching to the test." In such a case, any achievement gains merely reflect better test preparation and not long-term learning or "real" human capital.[21] To address these questions, I use data from a telephone survey that was conducted by the Ministry of Education among a sample of the English and math teachers who participated in the program.[22] For comparison purposes, a similar survey was conducted with a similar number of nonparticipating English and math teachers from the nontreated eligible schools and from other schools who on average had a matriculation rate equal to that of the treated schools. Therefore, the sample of schools included in this analysis does not overlap perfectly with the RT, RD, or the eligible schools samples. Since the number of interviewed teachers from RT sample schools is rather small, I conduct the analysis in this section with three samples, one with teachers from RT schools, a second with teachers from eligible schools, and a sample that includes all the interviewed teachers. As will be shown below, results from the three samples are generally consistent and indicate the same direction, yet they should be regarded as only suggestive.

Table A3 in the Web Appendix presents balancing tests that compare characteristics of treated and untreated teachers in each of these three samples for each subject separately. The evidence presented in this table suggests that the teachers in both subjects in each of the three samples had, on average, very similar characteristics. There are 13 variables in each column and in all columns there are at most two significant treatment-control differences, except for math teachers in the overall sample (column 2) where three of the 13 differences are significant. In the full and the RT English teacher samples, there are treatment-control imbalances

[21] See, for example, Paul Glewwe, Ilias Nauman, and Michael Kremer (2003).

[22] It is possible that teachers were aware that the survey was part of the incentives experiment and this may have affected their responses to these questions. To minimize such a "Hawthorne" type bias, the survey was presented to interviewees as a Ministry of Education general survey about matriculation exams and results, and the questions about the incentive program were placed at the end of the questionnaire.

TABLE 8— THE EFFECT OF PAY FOR PERFORMANCE ON TEACHING METHODS AND TEACHER EFFORT

	Math teachers					
	All interviewed teachers		Eligible schools' teachers		RT schools' teachers	
	Sample mean (1)	Treatment-control difference (2)	Sample mean (3)	Treatment-control difference (4)	Sample mean (5)	Treatment-control difference (6)
Teaching methods						
Teaching in small groups	0.661	0.007 (0.051)	0.557	0.111 (0.068)	0.525	0.193 (0.078)
Individualized instruction	0.614	−0.028 (0.060)	0.600	−0.014 (0.087)	0.600	−0.008 (0.125)
Tracking by ability	0.397	0.130 (0.059)	0.471	0.055 (0.073)	0.500	−0.035 (0.102)
Adapting teaching methods to student's ability	0.942	0.011 0.023	0.914	0.038 0.037	0.900	0.030 0.055
Teacher effort						
Added instruction time during the whole year, or before Bagrut exam	0.831	0.015 (0.036)	0.871	−0.025 (0.045)	0.825	−0.022 (0.072)
Added instruction time only before Bagrut exam	0.296	0.071 (0.048)	0.300	0.067 (0.069)	0.150	0.160 (0.082)
Number of additional weekly instruction hours	2.038	1.987 (0.600)	2.959	1.066 (0.809)	2.382	1.246 (0.965)
The teacher initiated the addition of instruction hours	0.709	−0.017 (0.051)	0.714	−0.022 (0.074)	0.625	0.093 (0.105)
Teacher's additional effort was targeted at						
All students	0.587	−0.025 (0.059)	0.614	−0.052 (0.079)	0.575	−0.012 (0.106)
Weak students	0.212	−0.058 (0.042)	0.214	−0.060 (0.059)	0.200	−0.045 (0.066)
Average students	0.011	0.043 (0.018)	0.029	0.025 (0.025)	0.025	0.017 (0.033)
Strong students	0.000	0.006 (0.006)	0.000	0.006 (0.006)	0.000	0.000 (0.000)
Number of teachers		358		239		111
Number of schools		109		68		27
Number of treated schools		46		46		17

in the proportion of teachers born in Israel, but this variable is balanced in the eligible schools sample. As will be shown below, the results across these three samples are very similar and so one can conclude that they are not derived by the differences in the proportion of teachers born in Israel.

Table 8 presents evidence concerning the effect of the incentive program on three behavioral outcomes of participating teachers: teaching methods, teacher effort, and focus of effort on weak or strong students. To help interpret the evidence, I should note that preparation for the matriculation exams at the end of twelfth grade is the essence and the focus of the curriculum of studies during the senior year in high school. Furthermore, high school seniors and their teachers end their regular school year in mid-March and spend the rest of the school year preparing for the

TABLE 8—THE EFFECT OF PAY FOR PERFORMANCE ON TEACHING METHODS AND TEACHER EFFORT *(continued)*

	English teachers					
	Sample mean (7)	Treatment-control difference (8)	Sample mean (9)	Treatment-control difference (10)	Sample mean (11)	Treatment-control difference (12)
Teaching methods						
Teaching in small groups	0.631	0.085 (0.054)	0.574	0.143 (0.073)	0.467	0.175 (0.127)
Individualized instruction	0.583	0.112 (0.060)	0.574	0.122 (0.078)	0.633	0.103 (0.109)
Tracking by ability	0.417	0.221 (0.058)	0.471	0.168 (0.087)	0.367	0.256 (0.113)
Adapting teaching methods to students ability	0.925	0.068 0.021	0.956	0.037 0.030	0.933	0.067 0.057
Teacher's effort						
Added instruction time during the whole year, or before Bagrut exam	0.564	0.223 (0.054)	0.544	0.243 (0.068)	0.567	0.207 (0.098)
Added instruction time only before Bagrut exam	0.207	0.211 (0.057)	0.147	0.271 (0.069)	0.167	0.192 (0.102)
Number of additional weekly instruction hours	1.144	1.458 (0.440)	1.655	0.946 (0.767)	1.040	1.148 (0.451)
The teacher initiated the addition of instruction hours	0.463	0.098 (0.064)	0.456	0.104 (0.082)	0.533	0.089 (0.101)
Teacher's additional effort was targeted at						
All students	0.330	−0.004 (0.064)	0.338	−0.012 (0.086)	0.433	−0.131 (0.098)
Weak students	0.197	0.129 (0.054)	0.176	0.150 (0.068)	0.133	0.225 (0.083)
Average students	0.016	0.020 (0.025)	0.015	0.021 (0.027)	0.000	0.057 (0.055)
Strong students	0.000	0.028 (0.013)	0.000	0.028 (0.013)	0.000	0.000 (0.000)
Number of teachers		329		209		83
Number of schools		105		64		25
Number of treated schools		42		42		15

Note: Standard errors in parentheses are clustered at the school level.

matriculation exams in various ways. Special marathon learning weekends away from school, for example, are very common.

The results, shown for English and math teachers separately, point to two patterns: the program modified teaching methods and led to an increase in teacher effort, as expressed in overtime devoted to student instruction after the regular school day. Added after-school instruction time was also observed among nonparticipating teachers, but was more prevalent among participating teachers. However, these effects are much more visible, more precisely measured, and relatively consistent across all three English teachers' samples.

The proportion of program-participant English teachers that taught in small groups, used individualized instruction and ability tracking, and adapted teaching methods to students' ability is higher than the respective proportions among nonparticipants. Most dramatic and significant is the difference between participating and nonparticipating teachers with respect to the proportion of teachers who use tracking by ability in the classroom: for example, in English, 64 percent versus 47 percent in the eligible sample and 62 percent versus 37 percent in the RT sample. Among math teachers, the only significant difference in teaching methods is in the prevalence of teaching in small groups: in the RT sample this practice was used by 71.8 percent of program math teachers as against 52.5 percent of comparison-group teachers, and the respective results in the eligible sample are similar.

Just over half of English teachers in the comparison group, as against 77 percent of participating teachers in the RT sample, reported that they added special instruction time throughout the school year. Among math teachers, over 80 percent of participating and nonparticipating teachers added instruction time beyond their regular teaching load. However, the answer to the question about added instruction time only during the exam preparation period, from mid-March to the end of June, reveals significant treatment-control differences both in math and English. Focusing on the RT sample results, the difference is very large, at 19.2 percentage points (16.7 percent versus 35.9 percent) among English teachers, and 16.0 percentage points (15.0 percent versus 31.0 percent) among math teachers. These differences are significantly different from zero in both subjects. A significant difference was found among English teachers in terms of the amount of instruction time added, just over one hour a week. A similar treatment effect is also evident among math teachers in all three samples although it is measured precisely only in the eligible sample.[23] Table 8 also reveals some differences in the targeting of effort among participating English teachers: 22.5 percent of participating teachers in the RT sample gave more attention to weak students as compared to only 13.3 percent among nonparticipating teachers. No such differences are found among the math teachers.

Beyond showing that the program induced changes in effort and pedagogy, this evidence is important because it indicates that the program enhanced forms of teaching and effort that teachers already practiced widely before the program started. This pattern greatly reduces the likelihood that the improvement in math and English matriculation outcomes reported above are traceable to new "teaching to the test" techniques that are less reflective of human capital accumulation. However, I should note that this is just a one-year experiment, and had the experiment lasted for two or three years, teachers might have adapted other, more efficient (from their point of view), methods of improving their chances in the tournament.

The results presented so far in the paper indicate that individual teachers matter in improving student outcomes and teacher productivity. Can one predict who the better teachers will be by some conventional measure of teacher quality? The correlation between the teachers' ranking in the tournament and their attributes may be used to characterize the good teachers, even though there is some noise in the ranking of teachers due to stochastic components in the performance of students. Estimates of such correlations for math and English teachers support the view that teaching quality is not highly correlated with characteristics such as age, gender, education, teaching certification, and years of teaching experience (Eric Hanushek 2002). None of these variables was very significant in explaining the ranking of teachers in the tournament. None of the teachers' attributes was significantly correlated with any measure of teacher effort discussed in the previous section. Other variables, however, showed significant correlations in the

[23] It should be noted that the teachers were informed that the program is a three-year experiment and therefore the possibility that the changes in effort reflected simply intertemporal substitution of effort this year and next cannot be overlooked.

regressions. Being born and educated outside of Israel had a positive influence on English teachers' effectiveness. Among English teachers educated in Israel, those who attended universities with the best reputations (Hebrew University of Jerusalem and Tel Aviv University) were significantly more effective than those who attended other universities or teachers' colleges. Among math teachers, the only attribute that had a significant effect on teaching effectiveness was mother's schooling: teachers whose mothers had completed high school or had earned a higher academic degree were much more effective than other teachers. No similar effect was found for father's education.

IV. Do Incentives Affect Teachers' Grading Ethics?

Incentive schemes may induce behavior distortions as agents seek to game the rules (see, for instance, Holmström and Milgrom 1991). Brian Jacob and Steven Levitt (2003), summarize the evidence on teachers' manipulation of test scores in programs that enhance accountability, and also provide evidence on outright cheating on the part of teachers and administrators who inflated test scores in various ways. The pay-for-performance scheme used in this experiment may have produced an incentive for teachers to inflate the school score, as the final matriculation score is a weighted average of the school score and the score in the national matriculation exam and the outcomes for ranking teachers were based on the final (weighted) scores. This is particularly important as teachers grade their own matriculation school exams. However, the bonus program included an explicit stipulation about sanctions that would apply to teachers who, according to Ministry of Education standards, would experience large gaps between the school scores and national scores. To assure comparability between school exam and national exam scores, the Ministry of Education has been using a scheme since 1996 (called "differential weighting") that includes supervision of the gaps between the school and the national scores and a set of rules about sanctions against schools when large gaps are found.[24] The stipulation in the bonus experiment determined that teachers who violate the differential weighting rules would be disqualified from the bonus program and consequently would not be eligible for bonus payments. However, the differential weighting scheme leaves much room for teachers to manipulate the school scores and still not violate the rules. For example, only an average gap of 20 points or more between the school and national score is considered an outlier (see details in footnote 24) and so teachers have enough room to inflate their school scores without violating the differential weighting rules.

In this section I present evidence that shows that teachers did not manipulate or inflate the school scores as a result of the program. The empirical evidence is based on a comparison of the discrepancies between the school and the national score in each exam and a comparison of these gaps between treated and control schools, while contrasting the respective evidence from the preprogram (2000) and program year (2001). The comparison of the two scores can be viewed as a natural experiment since the score in the national exam is an objective, unbiased measure of the student's knowledge while the school score may be biased due to teachers' cheating or other forms of test score manipulation.

Table 9 presents results for three samples: the RT sample (first panel), the RD sample (second panel), and the eligible schools sample (third panel). As noted above, the evidence pertains to

[24] A Ministry of Education document describes the rules of the differential weighting scheme in detail. If the *average* school score in a given exam is higher than the *average* national score by 20 points or more, or if it is lower by 10 points or more, the case is considered an outlier. If the probability of such an event is 1:10,000, the weights of the two scores are adjusted to 30 percent and 70 percent, respectively, instead of 50 percent each. If the probability of such an event is 1:1,000,000, the two scores are weighted at 10 percent and 90 percent, respectively. If outliers are defined in 8 percent or more of the exams, in at least two subjects, and in two of three consecutive years, disciplinary actions are taken against the school.

TABLE 9—ESTIMATES OF THE EFFECT OF INCENTIVES PROGRAM ON GRADING ETHICS

	Math					
	2000			2001		
	Treatment (1)	Control (2)	Diff. (3)	Treatment (4)	Control (5)	Diff. (6)
The randomized treatment sample						
School-state	0.096	0.159	−0.063	0.094	0.021	0.072
Score diff.	(0.068)	(0.046)	(0.080)	(0.079)	(0.049)	(0.092)
Observations	6,646	6,110	12,756	6,606	5,288	11,894
The rergression discontinuity sample						
School-state	−0.004	0.113	−0.117	−0.038	0.168	−0.206
Score diff.	(0.061)	(0.082)	(0.100)	(0.067)	(0.085)	(0.106)
Observations	6,374	3,912	10,286	6,130	3,248	9,378
Eligible schools						
School-state	−0.017	0.056	−0.073	−0.040	0.032	−0.072
Score diff.	(0.047)	(0.035)	(0.058)	(0.045)	(0.036)	(0.057)
Observations	14,414	16,130	30,544	13,768	14,102	27,870
	English					
	Treatment (7)	Control (8)	Diff. (9)	Treatment (10)	Control (11)	Diff. (12)
The randomized treatment sample						
School-state	−0.013	0.247	−0.260	0.026	0.047	−0.021
Score diff.	(0.093)	(0.204)	(0.221)	(0.076)	(0.081)	(0.110)
Observations	5,368	4,714	10,082	4,840	3,952	8,792
The regression discontinuity sample						
School-state	−0.154	0.122	−0.277	−0.099	0.031	−0.130
Score diff.	(0.060)	(0.119)	(0.130)	(0.060)	(0.070)	(0.090)
Observations	4,672	3,080	7,752	4,264	2,720	6,984
Eligible schools						
School-state	−0.066	0.052	−0.118	−0.097	0.040	−0.137
Score diff.	(0.055)	(0.052)	(0.075)	(0.055)	(0.037)	(0.066)
Observations	12,546	12,286	24,832	11,326	10,006	21,332

Notes: Standard errors in parentheses are adjusted for school-level clustering. Each of the entries in the table is the estimated difference between the school matriculation score and the score in the state matriculation exam. The entries in the columns noted "Diff." are the difference between the respective treated and controlled mean differences presented in the previous two columns in the same row.

the school year in which the bonus program was implemented, 2001, and also to the preprogram year, 2000, which will allow for difference-in-differences comparison. Each of the estimates in the table measures the difference between the school score and the score in the national exam. The scores were standardized to a distribution with zero mean and a unit standard deviation. Standardization was done separately for the school and national scores.

Of the six estimated differences in math in all three treated samples in 2001 and in 2000, four are negative and two are positive, but none of these differences is statistically different from zero. The respective differences estimated from the three control samples are all positive but only one of them is statistically different from zero. All but one of the difference-in-differences between treatment and control school means in a given year are negative, suggesting that teachers in treated schools "underestimate" on average their students' cognitive ability in math relative to their performance in the national exams and relative to students from the control schools. However, all these difference-in-differences estimates are not statistically different from zero except in one case (column 6, RD sample) and no systematic pattern is seen for the 2000 and

2001 difference-in-differences estimates. The results from triple differencing based on both years (column 6 minus column 3) strengthen this result.

The evidence for the English teachers is very similar to those of the math teachers. Therefore, based on the evidence presented in Table 9, it can be concluded that the performance incentive scheme discussed in this paper did not lead teachers to inflate artificially their students' school matriculation scores relative to the state matriculation scores in math and English. This could be the outcome of the sanctions implied by the Ministry differential weighting rules, or the threat of being disqualified from the tournament or of teachers having in general high ethical grading standards. However, the similarity of the evidence in the pre- and postprogram period suggests that the impartiality in grading was also evident before the bonus program.

V. The Israeli Experiment in a Broader Policy Context

The growing interest in incentive programs for teachers in the United States, Europe, and elsewhere has led to many new interventions. Pay-for-performance programs for US teachers include Minnesota's Q-Comp[25] $86 million merit pay initiative; Denver's Pro-Comp[26] $25 million teachers' pay-for-performance plan (Nancy Mitchell 2005); and Forida's E-Comp and STAR[27] programs. Chicago's public school system received a $27.5 million federal grant in 2006 to pilot a merit pay initiative for teachers, Chicago being the largest district in the country to experiment with performance-based pay.[28] An interesting and somewhat different program is the Dallas (Texas) Advanced Placement Incentive Program (APIP), which includes financial incentives for both teachers and students for each passing score earned on an AP exam in mathematics, science, and English (Kirabo C. Jackson 2007).[29] A recent *New York Times* article illustrated this trend, commenting that "A consensus is building across the political spectrum that rewarding teachers with bonuses or raises for improving student achievement, … can energize veteran teachers and attract bright rookies to the profession."[30]

Pay for teacher performance has also been implemented in other countries, for example, the United Kingdom's Pay Performance and Management Reform in 2000 (Adele Atkinson et al. 2004), the Victorian Government Schools Agreement 2001 in Australia,[31] Mexico's Carrera Magisterial Program, and Chile's SNED (Emiliana Vegas 2005). Smaller-scale randomized experiments have been implemented in India (Esther Duflo and Hanna Rema 2005; Venkatesh Sundararaman and Karthik Muralidharan 2008); and in Kenya (Glewwe, Ilias, and Kremer 2003).

Many of these programs include unique features, perhaps reflecting an uncoordinated effort to identify optimal incentive structures in the individual education systems. Similarly, the experiment studied in this paper has features not replicated elsewhere. Nevertheless, the findings reported here should be of broader interest, since the Israeli program has much in common with

[25] Q-Comp was proposed by Governor Tim Pawlenty and was enacted in Minnesota by the Legislature in July 2005. See http://education.state.mn.us/MDE/Teacher_Support/QComp/index.html.

[26] See Community Training and Assistance Center (2004) for an analysis of the Denver Pro-Comp pilot.

[27] For details of E-Comp, see http://www.fldoe.org/news/2006/2006_04_05/ValueTable.pdf and for STAR at http://www.fldoe.org/PerformancePay/.

[28] Other recent interventions of this type include programs in Mobile, Toledo, Columbus, Houston, Charlotte-Mecklenburg, and Dallas, and scores of other teacher performance-pay experiments are under way nationwide. Many of these programs are financed by federal grant funding (*s*ource: Department of Education).

[29] For more information on this program, visit http://www.collegeboard.com/student/testing/ap/about.html.

[30] Sam Dillon, "Long Reviled, Merit Pay Gains Among Teachers," *New York Times,* June 18, 2007 (http://select.nytimes.com/gst/abstract.html?res = F10715FA395B0C7B8DDDAF0894DF404482).

[31] This agreement was endorsed in December 2001 by 78 percent of the teachers who voted. Under the agreement, all teaching promotions are "linked to improvements in student learning," to be monitored via statewide testing of students in Math and English in years three and five. See http://www.wsws.org/articles/2001/jan2001/edu-j11.shtml.

performance-pay initiatives being tried elsewhere. For example, similar to recent experiments in the United States, the Israeli program relies on student achievement as the key benchmark, as opposed to any other measure of teacher performance such as professional standards of knowledge and training. Like Denver's ProComp program, Florida's E-Comp program and Dallas APIP programs, the Israeli program focuses on core academic subjects. The student outcomes were not based on special tests devised for the purpose of the program, but instead were based on state tests that are currently administered to measure academic achievement and that meet validity and reliability criteria. The Chattanooga program similarly relies on Tennessee's value added system to assess students' learning gains,[32] E-Comp uses Florida's Comprehensive Assessment Test (FCAT) system to measure students' performance, the Australian program uses an ongoing Achievement Improvement Monitor (AIM) system that includes testing students in math and English,[33] and the UK program uses test scores from an ongoing national program. In the Israeli program, student outcomes were adjusted for contributing factors, such as student socioeconomic characteristics, as is common in other programs.[34]

The Israeli program used a well-known method to select the teachers to be awarded. As in Florida's E-Comp rank order tournament framework, math and reading teachers in the Israeli program were ranked according to points they received based on the progress of their students, so that bonuses were based on improvements in student learning. Similar to the Florida program, more points were assigned to outcomes that are more highly valued and less likely to be achieved.[35] Also, like in most other programs, the performance awards were one-time bonuses and thus were not added to base pay. The average bonus awards were relatively generous, similar to some programs in the United States that have given bonuses up to $20,000 (Allan Odden and Marc Wallace 2007). For example, the APIP can deliver a considerable increase in compensation and indeed some teachers gained more than $11,000 in annual earnings (Jackson 2007). As noted in Section III, the Israeli program included multiple levels of bonuses as in other programs in the United States.[36] Another common feature of the Israeli and other programs is that all teachers were eligible for the bonuses offered, although in practice only a subset of teachers were rewarded. This feature strikes a desirable balance, since if too many teachers are rewarded, teachers may not need to invest much effort to benefit from the program.

The unique features of the Israeli policy experiment are especially interesting and may be a model for others. For example, the Israeli approach relies on a comparison of test scores in internal and external tests and imposes severe sanctions for major gaps between them, as explained in Section V above. This reduces the possibility of cheating of the sort discussed by Jacob and Levitt (2003) in Chicago's public schools. The Israeli program also leaves most decisions about

[32] See http://mb2.ecs.org/reports/Report.aspx?id = 1131 and Claire Handley and Robert A. Kronley (2006), for details of this program, which used measures of student progress to identify teachers qualifying for a wage premium.

[33] The Australian Achievement Improvement Monitor (AIM) system includes testing students in math and English in grades 3 and 5. In 2001 this was extended to grade 7, the first year of high school.

[34] North Carolina has operated a similar program for over five years (Kelley, Heneman, and Milanowski 2002). An alternative approach is to compare the absolute gain relative to a predetermined standard, as in Kentucky in the 1990s (Kelley, Conley, and Kimball 2000).

[35] The centerpiece of E-Comp required all school districts in Florida to identify the top 10 percent of teachers in each field and award them a 5 percent salary supplement. Those who teach math and reading are ranked exclusively according to how much their students have improved their scores over the previous year. Teachers earned points when they advanced their students from one level of proficiency to another. In 2006 Florida replaced E-Comp with a similar performance pay plan called Special Teachers Are Rewarded (STAR), with an annual budget of $147.5 million. This plan allowed districts and charter schools to implement performance pay provisions of section 1012.22, Florida Statutes, and to access their portion of the funds in the appropriation. For details, see http://www.fldoe.org/PerformancePay/pdfs/STAR_SuptMemo.pdf.

[36] The Israeli program had four levels of bonuses, and the Cincinnati and Charlotte-Mecklenburg program, for example, had two levels of awards (Odden and Wallace 2007).

how to attain performance targets up to the teachers. This increased flexibility appears to have paid off. Another aspect of the program that was shown to be very effective here, and therefore lessons might be drawn from it for other designs, is the use of multiple outcomes to measure teacher performance and to signal which gains are more socially desirable. The evidence shown here indicates that teachers responded to the specifics of the incentive scheme, especially to features that signaled different rewards for gains in different outcomes. The general lesson from this result is that strategic monetary signals in an incentive scheme in schools can induce teachers to target their effort in a direction desirable to policymakers. It also can make the intervention more effective.

VI. Conclusions

The evidence presented in this paper indicates that pay-for-performance incentives can align the interests of schoolteachers with the interests of the school system without necessarily inducing behavior distortions such as test score manipulations or teaching-to-test practices. This result is evident despite the widely held concern about the team nature of learning in school, i.e., the belief that a student's results are the outcome not of the inputs of a single teacher but of the joint contributions of many teachers. The magnitude of the estimated effects and the evidence about teachers' differential efforts under an incentive regime suggest that teacher incentives should be considered as a promising method of improving school quality.[37] These results about individual teacher incentives and the earlier evidence about the effect of group school incentives (Lavy 2002) are important in the policy context of teacher compensation and schooling quality. However, the caveat of the results presented in this paper is that the experiment lasted for just one year and, therefore, it does not permit us to study the effects of the incentives on other cohorts and to identify long-term effects. For example, beyond affecting motivation, teacher incentives may also have a long-term effect by means of the screening and selection of teachers (Edward Lazear 2003). They may also have possible dysfunctional or counterproductive long-run responses. Estimating such potential positive and negative long-term effects is not feasible in this study and should be a subject of future research.

The nonrandom nature of the assignment of schools and teachers to the experiment entailed alternative identification strategies. The natural experiment that resulted from the measurement error in the assignment variable provided an appealing approach to the problem of nonrandom assignment into the incentive pay program. However, the very close similarity between the RT and the RD results suggests that the regression discontinuity method worked as well in this case, while the difference-in-difference estimates based on the full sample of eligible schools produced biased estimates.

I have shown that individual teachers' incentives worked their effect through two channels: causing more students to take a matriculation exam than otherwise would have, and increasing the pass rate and the mean test score among students who would have taken the exam regardless of the program. Even though the first outcome may seem easier to manipulate by teachers, the fact that both the unconditional and the conditional pass rate and mean test score increased in both subjects suggests that the program also improved the quality of students' knowledge and skills among those who took the exam because of the program. This can be seen by the fact that more than half of the increase in the unconditional pass rate in math and about a fourth of its increase in English is due to the higher test taking rate.

[37] Victor Lavy and Analia Schlosser (2005) provide compelling evidence about the costs and benefits from the teachers' incentive intervention, both relative to other interventions and also by comparing the program cost per student to the likely economic benefits of the improved outcomes.

On the basis of a postprogram survey among participating and nonparticipating teachers, I found evidence that links the improvement in students' cognitive outcomes on math and English matriculation exams to changes in participating teachers' teaching methods, pedagogical techniques, and additional effort during the program. Teaching in smaller groups and tracking students by ability, for example, seemed much more prevalent among participating teachers, who also enhanced a practice that is very common among all teachers, i.e., adding additional teaching time during the four-month period in which they prepare students for the matriculation exams.

The structure of the Israeli matriculation exam system, which is based on compulsory testing at the end of high school and a minimum number of required credits, closely resembles the corresponding systems used in France, Germany, Italy, New York, Massachusetts, and other locations. The structure of the Israeli teachers' incentive program also had much in common with performance-pay initiatives being tried in the United States and elsewhere. As a result of these similarities, the results and lessons drawn from the experiment examined in this paper are relevant for many education systems in Europe and the US.

REFERENCES

Atkinson, Adele, Simon Burgess, Bronwyn Croxson, Paul Gregg, Carol Propper, Helen Slater, and Deborah Wilson. 2004. "Evaluating the Impact of Performance-related Pay for Teachers in England." University of Bristol Centre for Market and Public Organisation Discussion Paper 04/113.

Community Training and Assistance Center. 2004. Catalyst for Change, Boston.

Duflo, Esther, and Rema Hanna. 2005. "Monitoring Works: Getting Teachers to Come to School." National Bureau of Economic Research Working Paper 11880.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2003. "Teacher Incentives." National Bureau of Economic Research Working Paper 9671.

Green, Jerry R., and Nancy L. Stokey. 1983. "A Comparison of Tournaments and Contracts." *Journal of Political Economy*, 91(3): 349–64.

Handley, Claire, and Robert A. Kronley. 2006. *Challenging Myths: The Benwood Initiative and Education Reform in Hamilton County.* Atlanta: Kronley & Associates.

Hanushek, Eric A. 2002. "Publicly Provided Education." National Bureau of Economic Research Working Paper 8799.

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

Israel Ministry of Education, High School Division. 2000. "Individual Teacher Bonuses Based on Student Performance: Pilot Program." December, Jerusalem (Hebrew).

Israel Ministry of Education. 2001. *Statistics of the Matriculation Examination* (*Bagrut*) *Test Data, 2000.* Jerusalem: Ministry of Education Chief Scientist's Office.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*, 118(3): 843–77.

Kirabo, Jackson C. 2007. "A Little Now for a Lot Later: A Look at a Texas Advanced Placement Incentive Program." Unpublished.

Kelley, Caroline, Herbert Heneman, and Anthony Milanowski. 2002. "Teacher Motivation and School-Based Performance Awards." *Education Administration Quarterly* 38(3): 372–401.

Kelley, Caroline, S. Conley, and S. Kimball. 2000. "Payment for Results: The Effects of the Kentucky and Maryland Group-based Performance Award Programs." *Peabody Journal of Education* 75(4): 159–99.

Lavy, Victor. 2002. "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement." *Journal of Political Economy*, 110(6): 1286–317.

Lavy, Victor. 2004. "Performance Pay and Teachers' Effort, Productivity and Grading Ethics." National Bureau of Economic Research Working Paper 10622.

Lavy, Victor. 2007. "Using Performance-Based Pay to Improve the Quality of Teachers." *The Future of Children*, Spring: 87–110.

Lavy, Victor, and Analía Schlosser. 2005. "Targeted Remedial Education for Underperforming Teenagers: Costs and Benefits." *Journal of Labor Economics*, 23(4): 839–74.

Lazear, Edward P., and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy*, 89(5): 841–64.

Lazear, Edward. 2003. "Teacher Incentives." *Swedish Economic Policy Review*, 10(2): 179–214.

Neal, Derek, and Diane Whitmore Schanzenbach. 2007. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." National Bureau of Economic Research Working Paper 13293.

Odden, Allan, and Marc Wallace. 2007. "Rewarding Teacher Excellence." Unpublished.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature*, 37(1): 7–63.

Sadowski, Connie. 2006 "Houston District OKs Teacher Merit Pay Plan." The Heartland Institute, March, http://www.heartland.org/policybot/results/18637/Houston_District_OKs_Teacher_Merit_Pay_Plan.html.

Venkatesh, Sundararaman, and Karthik Muralidharan. 2008. "Teacher Performance Pay: Experimental Evidence from India." Paper presented at the Ninth Neemrana Conference on the Indian Economy, Neemrana, India.

Wakelyn, David J. 1996. "The Politics of Compensation Reform: A Colorado Case Study." Paper presented at the 20th Annual American Educational Finance Association Conference, Salt Lake City, UT.

Vegas, Emiliana, ed. 2005. *Incentives to Improve Teaching: Lessons from Latin America*. Washington, DC: World Bank.

This article has been cited by:

1. VICTOR LAVY. 2010. Effects of Free Choice Among Public Schools. *Review of Economic Studies* **77**:3, 1164-1191. [CrossRef]
2. C. Kirabo Jackson. 2010. Do Students Benefit from Attending Better Schools? Evidence from Rule-based Student Assignments in Trinidad and Tobago*. *The Economic Journal* no-no. [CrossRef]