

**Paying for Performance: The Effect of Individual Financial Incentives
on Teachers' Productivity and Students' Scholastic Outcomes***

Victor Lavy

The Hebrew University of Jerusalem and CEPR

July 2003

* Special thanks go to Alex Levkov for his outstanding research assistance. I also thank Josh Angrist, Abihijit Banerjee, Eric Battistin, Esther Duflo, Caroline M. Hoxby, Andrea Ichino, Hessel Oosterbeek, Yona Rubinstein, and seminar participants at EUI, MIT, Hebrew, Princeton and Tel Aviv University, The Tinbergen Institute, the NBER 2002 Summer Institute and the 2002 UK Public Economics Conference for their helpful discussions and comments. The 2001 Teachers' Incentive program was funded by the Israel Ministry of Education and administered by the Teaching Staff Division. The views expressed in this paper are those of the author alone and have not been endorsed by the program sponsors.

Paying for Performance: The Effect of Individual Financial Incentives on Teachers' Effort and Students' Scholastic Outcomes

Abstract

Performance-related incentive pay for teachers is being introduced in many countries, but there is little evidence of its effects. This paper evaluates a rank-order tournament among teachers of English, Hebrew, and mathematics in Israel. Teachers were rewarded with cash bonuses for improving their students' performance on high-school matriculation exams. Two identification strategies were used to estimate the program effects, a regression discontinuity design and propensity score matching. The regression discontinuity method exploits both a natural experiment stemming from measurement error in the assignment variable and a sharp discontinuity in the assignment-to-treatment variable. The results suggest that performance incentives have a significant effect on directly affected students with some minor spillover effects on untreated subjects. The improvements appear to derive from changes in teaching methods, after-school teaching, and increased responsiveness to students' needs. The program appears to have been more cost-effective than school-group cash bonuses or extra instruction time and is as effective as cash bonuses for students.

Victor Lavy
The Hebrew University of Jerusalem, Department of Economics
msvictor@mscc.huji.ac.il

1. Introduction

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 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. Several dilemmas and challenges arise in the task of designing and evaluating teachers' performance incentives. How should teacher performance be measured? How can individual teachers' contributions be identified? How should the rewards be structured and how generous should they be? Do teachers' pedagogy and effort respond to financial incentives? Are teachers' performance incentives more effective than school-based performance rewards? How relevant and important are the spillover or substitution effects of teachers' incentives? The evidence presented in this paper relates directly to these questions and is based on results of a pay-for-performance experiment among a sample of high-school teachers in Israel.

This paper evaluates an Israeli program that rewarded teachers with cash bonuses for improvements in their students' performance on the high-school matriculation exams in English, Hebrew, and mathematics. The bonus program was structured in the form of a rank-order tournament among teachers, in each subject separately.² Thus, teachers were rewarded on the basis of their performance relative to other teachers of the same subjects. Relative performance was preferred over measurements based on absolute performance for two reasons: these awards would stay within budget and there were no obvious standards that could be used as a basis for absolute performance measures. The relative measurements were based on comparison of the achievements of each teacher's students with predicted values using regressions. Two measurements of students' achievements were used as indicators of teachers' performance: the passing rate and the average score on each matriculation exam. The total amount to be awarded in each tournament was predetermined and individual awards were determined on the basis of rank and a predetermined award scale.

¹ Examples include performance-pay plans in Dade County, Florida, Denver, Colorado, and Dallas, Texas, in the mid-1990s; statewide programs in Iowa and Arizona in 2002; programs in Cincinnati, Philadelphia, and Coventry (Rhode Island); and the Milken Foundation TAP program. In the UK, the government recently concluded an agreement with the main teachers' unions on a new teachers' performance-pay scheme starting in 2002/2003, with a budget of nearly £150 million. In New Zealand, the government completed a system-wide program of performance-related pay for teachers in 2001. For discussion and analysis of these programs, see Clotfeller and Ladd, 1996; Elmore, Abelman and Fuhrman, 1996; Kelley and Protsik, 1996.

The main questions of interest in this experiment relate to the effect of the program on teachers' pedagogy and effort and on the effect of the experiment on students' achievements. The paper attempts to answer the following key questions: did the program cause teachers to exert more effort, change their pedagogy, improve their preparation and teaching, and evaluate more effectively their students' need for additional instructional assistance? Did the students' outcomes improve as a result of the program? Did the program have spillover effects on students' outcomes in untreated subjects? How effective was the program relative to other relevant interventions?

Although the program was designed as an experiment, schools were not assigned to it at random. Therefore, the search for answers to the foregoing questions was complicated by the possibility that the schools included in the program were a selective sample with attributes that might be related to students' outcomes for reasons other than those related directly to the intervention.

Two alternative identification strategies were used to estimate the causal effect of the program. The first was a regression discontinuity (RD) design based on the assignment rule that determined program participation. This process was based on a threshold function of an observable assignment variable, the 1999 school matriculation rate: schools with this rate equal to or lower than a critical value (45 percent) were included in the program; others were excluded. This RD design may be described as $T = 1\{S \leq 45\}$, where T is an indicator of assignment to treatment and S is the assignment variable. I developed two empirical variants on the basis of this RD framework. The first was based on a measurement error in S (the variable used to assign schools to the program): $S = S^* + \varepsilon$, where S^* is the true rate and ε is a 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, T was randomly assigned, conditional on S^* . Since this random assignment obtained mostly for schools that were near the threshold, controlling for S^* , potentially in a fully non-parametric way, defined a *natural experiment* that may still be viewed as an RD strategy based on a covariate that has an element of random variation. This identification strategy was enhanced by the use of available panel data (before and after the program) that allowed an estimation of differences-in-differences estimates in the natural experiment setting.

A second variant on the RD design, which I used for identification, was based on the classic notion of an RD design, i.e., that the likelihood of an S value slightly above or below the threshold

² See Lazear and Rosen (1981), Green and Stokey (1983) and Prendergast (1999) for discussion of the theory of individual and group incentives in rank-order tournaments.

value of the assignment variable is largely random. If this is true, then treated and untreated schools in a narrow band around the threshold might be undistinguishable in potential outcome. However, a weaker assumption is based on controlling for parametric functions of S . In other words, conditional on X , we expect no variation in T . I exploited this sharp discontinuity feature in the assignment mechanism to define a second, more “conventional,” variation of an RD design to estimate the effect of the teachers’ incentive program. Here, as before, I exploited the panel nature of the data and embedded the sharp RD identification in a differences-in-differences estimation.

The second identification strategy that I used exploited the very rich and unique data available on all schools and students, including many measures of lagged outcomes, to build a comparison group by matching. The matching is based on the propensity score matching (PSM) method. The availability of various dimensions of lagged outcomes improved the likelihood of matching pupils in view of non-observable attributes as well as observable ones. The PSM results were compared with the results of traditional regression estimates, which may be viewed as a conventional baseline control strategy.

Section 2 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 3 discusses the evaluation strategy. Section 4 presents the PSM strategy and the results of its application in order to identify and estimate the causal effect of teachers’ incentives on the mathematics and English performance of students. Section 5 presents the two variants on the RD method and presents the respective empirical results. Section 6 presents evidence of the effect of incentives on teachers’ effort and pedagogy. Section 7 discusses the correlation between teacher attributes—such as quantity and quality of schooling, teaching experience, age, gender, and parental schooling—and performance in the tournament. Section 8 presents evidence of the relative effectiveness (cost-benefit) of paying teachers for performance and other interventions, such as school group incentive programs and monetary incentives for students.

The results suggest that incentives increase student achievements by increasing the attempt rate and the passing rate of exams. The improvement appears to come from changes in teaching methods, after-school teaching, and increased responsiveness to students’ needs. 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” that do not produce real learning.³ Finally, the cost-benefit comparison of other relevant interventions suggests that financial

³ On this point see for example, Glewwe, Ilias, and Kremer, 2003.

incentives for individual teachers are more efficient than teachers' group incentives and as efficient as paying students monetary bonuses to improve their performance. All three incentive programs were more efficient than a program that targeted instruction time to weak students.

2. Tournaments as a Performance Incentive

2.1 Theoretical Context

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 (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 and multitask jobs. 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 themselves 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.⁴

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 (Lazear and Rozen, 1982; Green and Stokey, 1983). The only question that matters in

winning such tournaments is how well one does relative to others and not the absolute level of performance.. Although promotion is not an important career feature among teachers, emphasize on relative rather than 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 where there are no obvious standards that may be used as a basis for absolute performance, relying on how well teachers do relative to others seems a preferred alternative. Therefore, we used the structure of a rank-order tournament for the teacher-incentive experiment described below.

2.2 Secondary Schooling in Israel

Lavy (2002) presents the results of a *group incentive* experiment in Israel (1995–1999), in which schools competed on the basis of their average performance and the rewards were distributed equally among all teachers in the winning schools. The purpose of the program was to improve students' achievements on the *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 (hereinafter: credits) per subject.⁵ 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. A minimum of twenty credits is required to qualify for a matriculation certificate. About 52 percent of high-school seniors received matriculation certificates in 1999 and 2000, i.e., passed enough exams to be awarded twenty credits by the time they graduated from high school or shortly thereafter (Israel Ministry of Education, 2001).

In early December 2000, the Ministry of Education unveiled a new teachers' bonus experiment in forty-nine Israeli high schools. 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 ten

⁴ See Jenson and Murphy, 1990; Holmstrom and Milgrom, 1991; Milgrom and Roberts, 1992; Gaynor and Pauly, 1990; Kandel and Lazear, 1992; Gibbons, 1998; Malcomson, 1998 and Prendergast, 1999; for a discussion of these issues in the general context of incentives.

⁵ In Israel, a high school matriculation certificate is a prerequisite for university admission and one of the most economically important education milestones. Many countries and some American states have similar high school matriculation systems. Examples include the French Baccalaureate, the German Certificate of Maturity (Reifezeugnis), the Italian Diploma di Maturità, the New York State Regents examinations, and the recently instituted Massachusetts Comprehensive Assessment System.

through twelve 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 forty-nine schools. The program was described to them as a voluntary three-year experiment.⁶ All the principals reacted very enthusiastically to the details of the program. One principal changed his mind later and removed his school from the program. A survey among all participating teachers showed us that 92 percent knew about the program and that 80 percent were familiar with the details of how the winners and the size of the bonuses would be determined.

Three formal rules guided the assignment of schools to the program: only comprehensive high schools (having 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). Ninety-seven schools met the first two criteria; forty-nine met the third one.⁸

Schools were also allowed to replace the language (Hebrew and Arabic) teachers with teachers of other core matriculation subjects (Bible, literature, or civics). Therefore, the evaluation may include English and math teachers only because school participation in Hebrew and Arabic was choice outcome. Since some schools did exercise this option, the sample of schools that elected not to do so is endogenous.

2.3 The Israeli Teacher-Incentive Experiment

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. Teachers were ranked in view of their classes' passing rate and mean score. 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, their level of proficiency in each

⁶ Due 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 will not continue as planned for a second and third year.

⁷ Performance was measured in terms of the average passing rate in the mathematics and English matriculation tests during the last four years (1996–1999). If any of these rates was 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 failing rate among matriculation subjects.

subject, and a fixed school-level effect. Separate regressions were used to compute the predicted passing rate and mean score, and each teacher was ranked twice, once for each outcome. The school submitted student enrollment lists that were itemized by grades, subjects, and teachers. The reference population was the enrollment on January 1, 2001, the starting date of the program. All students who appeared on these lists (including dropouts and students who did not take the June 2001 exams, irrespective of the reason) were included in the class mean outcomes at a score of zero.

All teachers who had a positive residual (actual outcome less predicted outcome) 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 passing 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.⁹

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. Three hundred and two teachers won awards—94 English teachers, 124 math teachers, 67 Hebrew and Arabic teachers, and 17 among the other subjects. Three English teachers won two awards each, twelve math teachers won two awards each, and one Hebrew teacher won two first-place awards totaling \$15,000.

We conducted a follow-up survey of teachers in the program during the summer vacation after the end of the school year. Seventy-four percent of teachers were interviewed. Very few of the intended interviewees were not interviewed, most of which due to wrong phone numbers or teachers who could not be reached by phone after several attempts. The survey results show that 92 percent of the teachers knew about the program, 80 percent had been briefed 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

⁸ A relatively large number of religious and Arab schools met all three selection rules. To keep their proportion in the sample close to their share in the population, the matriculation threshold for these schools was set to 43 percent.

⁹ For more details, see Ministry of Education, High School Division, “Individual Teacher Bonuses Based on Student Performance: Pilot Program,” December 2000, Jerusalem (Hebrew).

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 students' achievements.

2.4 The Data

The data I 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 micro student files included the full academic records of each student on 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. Each Bagrut exam is written at the Ministry of Education by an independent agency. There are two exam periods, winter (January) and summer (June), and all pupils are tested in a given subject at the same date. The exams are graded centrally; each exam by two independent external examiners, and the final score is the average of the two. This protocol eliminates the possibility of teachers grading their own students' exams and thereby reduces the possibility of cheating.

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.

I defined three outcomes for each subject: the number of tests taken by a student in the given subject,¹⁰ the total number of credits that passage of these tests confers, and the total credits earned. The second and third measurements reflect the proficiency level of the curriculum of each study program. Another important aspect of the evaluation is the overall effect of the program on the students' Bagrut certification. Certification—the accumulation of twenty credits or more—is a very important educational milestone that is highly rewarded in the labor market and is a necessary ticket to higher education. Estimates of the effect of the program on this outcome will be presented along with the subject specific outcomes.

Table 1 presents descriptive statistics for the 2000 and 2001 cohorts of high-school seniors for two samples, the forty-nine schools included in the program and all other high schools. The table reveals that the means of students' characteristics in treated schools differed from the corresponding

¹⁰ Each matriculation subject may involve more than one exam. For example, the mathematics curriculum includes two tests and the English program includes three—two written and one oral.

means in all other schools. Large differences between the sets of schools were also observed in the means of lagged students' outcomes and school characteristics. By implication, the sample of the forty-nine treated schools is not a representative sample of high schools in Israel.

3. OLS and Propensity Score Matching

Let us define the outcome of student i whose teacher participated in the incentive program as Y_i^1 and the outcome of a student whose teacher was not included in the program as Y_i^0 . Thus, the effect of the intervention on the i th pupil is $(Y_i^1 - Y_i^0)$ and it is not observed because either one or the other outcome is observed. The parameter of interest is the estimated effect of treatment on those treated, i.e., $E(Y_i^1 - Y_i^0 | T_i = 1)$, where T is 1 for students in schools with participating teachers and 0 for students of non-participating teachers. What is observed is $E(Y_i^1 | T_i = 1)$, the average outcome for students whose teachers participated in the incentive program. An OLS regression is the simplest model that may surmount this difficulty and help to construct the counterfactual $E(Y_i^0 | T_i = 1)$. The causal interpretation of the OLS estimate, which may serve as a benchmark with which we may compare the results of other models, is based on the assumption that we may account for all potential outcome differences between treated and untreated students by controlling for their observable characteristics ("selection on observable"). On the basis of this assumption, I estimate a model that includes as controls students' characteristics—including lagged outcomes—and school covariates, using a sample that includes all 12th-grade students in all high schools countrywide. The model may be expressed as:

$$(1) \quad Y_{ij} = \alpha_j + X_{ij}' \beta + Z_j' \gamma + \delta T_{ij} + \varepsilon_{ij}$$

where i indexes students; j indexes schools; T is the assigned treatment status, X is a vector of student-level covariates, and Z is the vector of school-level covariates. X and Z include, respectively, the entire individual and school-level variables presented in Table 1.

Table 2 presents the OLS estimates of the program effects on the math and English outcomes. The standard errors reported in the table are adjusted for clustering, using formulas set forth in Liang and Zeger (1986). The treatment-effect estimates in English and math are all positive and significantly different from zero. These results suggest that the program caused students to take more exams, attempt to earn more credits, and experience a higher passing rate and therefore earn more credits in math and English. The size of the effect along each of these channels is similar within each subject. In math, the program led to an 18 percent increase in the number of attempted exams and attempted credits, relative to the control-group mean of these two outcomes, and to a 14 percent increase in

credits earned relative to the respective control-group mean. In English, the effects were smaller—a 10 percent increase relative to the respective mean of the three outcomes—but again, the size of the effect was equal for all three outcomes. The estimated effect of the program on the matriculation rate (lowest row in Table 2) is an increase of 5.4 percent in the math sample and 4.2 percent in the English sample, both significantly different from zero.

An alternative to the simple model of a controlled regression is identification based on matching. Matching may be implemented non-parametrically by defining cells using discrete characteristics.¹¹ The more characteristics there are, however, the harder it becomes to find untreated individuals who are identical to treated individuals. Rosenbaum and Rubin (1985) suggest a solution to this dimensionality problem: a weighted index of each individual’s characteristics, referred to as “propensity score matching” (PSM).¹² The first empirical step in implementing this method is to estimate the propensity score for each student, using a regression of student and school characteristics on treatment status.¹³ The control sample is restricted to only those observations whose propensity-score value falls within the range of the propensity score in the treatment sample. By imposing this common support condition in the estimation of the propensity score, we improve the quality of the matches and avoid a major source of bias (Heckman, Ichimura, and Todd, 1997). Students are matched according to their propensity score by the *Nearest Neighbor Matching* method within 100 intervals.¹⁴ In estimating the treatment effect by the matching method, corrected standard errors are derived by using numerical bootstrapping methods.

Table 3 presents descriptive statistics for the treated and control groups in a PSM sample that was based on students enrolled in English and math classes separately. The English and math samples

¹¹ For an application of this method, see Angrist (1998).

¹² To construct the counterfactual $E(Y_i^0 | P_i = 1)$ within the propensity-score framework, the following assumption is needed: $E(Y_i^0 | T_i = 1, X_i, Z_j) = E(Y_i^0 | T_i = 0, X_i, Z_j)$ which means that given the observable characteristics of students (X_i) and schools (Z_j), the placement in treatment and control groups is random. Under this assumption, it is now well known (see Rosenbaum and Rubin, 1983) that $E(Y_i^0 | T_i = 1, \Pr(T_i = 1, | X_i, Z_j)) = E(Y_i^0 | T_i = 0, \Pr(T_i = 1, | X_i, Z_j))$ where $\Pr(T_i = 1, | X_i, Z_j)$ is the propensity score and is simply the probability of being assigned to treatment given observed characteristics. It follows that the counterfactual can be estimated by the sample analog of $E(Y_i^0 | T_i = 1) = E_{F^1}[E(Y_i^0 | T_i = 0, \Pr(T_i = 1, | X_i, Z_j))]$, where E_{F^1} denotes an expectation about the distribution of the propensity score in the treatment sample.

¹³ Dearden, Emmerson, Frayne and Meghir (2003) provide a recent example of the application of the propensity score matching approach in evaluating an education program.

¹⁴ Our unusually rich data, which include many lagged achievement outcomes, probably improves the match of important unobserved individual attributes such as ability and motivation.

were different, reflecting the difference in the number of students who took English and math during the experiment. All variables that appear in the table were used in the matching equation. Matches were found for almost all treated students (95 percent) in both the math and English samples, from 330 schools in the English sample and from 350 schools in the math sample.

The first panel in Table 3 shows that the matching process leads to a perfect match since none of treatment–control differences is statistically different from zero except for immigrant status. The second panel in the table shows that the two samples are also perfectly balanced in all six measures of lagged outcomes. These results reinforce our confidence that the two samples are also well balanced in terms of unobserved student covariates.¹⁵

Table 4 presents the results of estimating equation (1) using a sample that includes the treated students from the forty-nine schools that were included in the program and their matches as described above.¹⁶ The treatment-effect estimates presented in Table 4 are qualitatively very similar to the OLS program effect estimates presented in Table 2. Focusing for comparison on the effect of treatment on credits earned, the PSM estimate for the math sample is 0.293 while the OLS estimate is 0.228. The PSM estimate for English credits earned is 0.145 while the OLS estimate is 0.102. The estimated effect of the program on the matriculation rate, based on the PSM method and the math sample, is 0.051 (S.E.=0.013), again very similar to the OLS math sample estimate (0.054). The overall close similarity between the OLS estimates and the PSM estimates is an indication that a simple control for students’ characteristics and lagged outcomes and for school characteristics is quite sufficient in this case.

3.1 Allowing for Heterogeneity in the Effect of Treatment by Student Ability

As an additional check on the causal interpretation of the results presented in Table 2 and 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

¹⁵ As another check of the quality of matching, I re-estimated the propensity score model, omitting from the equation all lagged outcomes except the math and English lagged credits. I then checked how well balanced the treated sample and its comparison counterpart were in terms of the omitted lagged outcomes variables. . None of the mean differences of these outcomes (history and biology credits, total credits, average score) was significantly different from zero.,

¹⁶ The standard errors were estimated using bootstrapping techniques. To account for clustering in the error term, I used a procedure that included, in each round of estimation, a random draw of samples of students (treatment and control separately) and a random draw of schools (treatment and control separately). In terms of the asymptotic bias in the estimates of the standard errors, the matching on the propensity score has an advantage because of the relatively large number of clusters (schools).

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, I estimated the following model for each of the three outcome of interest in English and math:

$$(2) \quad Y_{ij} = \alpha + X_{ij}'\beta + Z_j' \gamma + \sum_q d_{qi} \mu_q + \sum_q \delta_q T_j + \varepsilon_{ij},$$

where δ_q is a quartile-specific treatment effect and μ_q is a quartile main effect. Students with very high 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 in 2000 was 90 percent. Therefore, one would not expect to find an effect of the teacher-incentive program on students in this quartile. In contrast, students with scores around or below the mean of the score distribution fell into a range in which extra effort—of their teachers and of themselves—may have made a difference. Therefore, I looked for significant estimates for students mainly in quartiles 1–3.

Table 5 reports results of the estimation of equation (2), which allows treatment to vary by quartile of the mean-score distribution in matriculation exams taken before the program, using the PSM sample of treated students (the same sample that was used in Table 4). The pattern in the table suggests that the average effects reported in Table 4 for all three outcomes originate in the effects on the first two quartiles—the below-average students—while no significant effects were estimated for above-average students (quartiles 3 and 4). The zero effect on these outcomes in the third and fourth quartiles is not surprising since all students in these quartiles were expected to take all exams as scheduled. This pattern was evident similarly in math and in English. There were few deviations from this overall pattern, most notably the math outcomes of attempted exams and credits, for which some positive effects were also evident for students in the third quartile.

Table 5 also presents results about the effect of the program on the matriculation rate by quartile of lagged achievements. The pattern is very similar to that of the other outcomes in Table 5: a significant effect for the two first quartiles (in the math sample, for example, increases of 4.1 percent and 10.4 percent for students in the first and second quartile, respectively) and no significant effect in the upper two quartiles.

4. Regression Discontinuity

4.1 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 matriculation-rate data given to administrators were culled from a preliminary and incomplete file of matriculation status. For many students, matriculation status was erroneous since it was based on missing or incorrect information. The Ministry later corrected this preliminary file, as it does every year.¹⁷ As a result, the matriculation rates used for assignment to the program were inaccurate in a majority of schools. The 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 by 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 1999.¹⁸ Most (80 percent) 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 co-move; their correlation coefficient is very low, at -0.085 , even though the p-value that it is different from zero is 0.055. However, if a few extreme values (five schools) are excluded, the correlation coefficient becomes basically zero. Although the figure may suggest 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. Similar evidence arises when the sample is limited to the 97 schools that were eligible for treatment, those from which 49 schools were assigned for treatment (Figures 3 and 4). If the two extreme values in Figure 4 are excluded from the sample, the estimated correlation coefficient between the correct 1999 matriculation rate and the measurement error rate, although negative, is practically zero. Similar evidence is 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 (Figure 4A).

A further check on the random nature of the measurement error can be based on its correlation with other student or school characteristics that might be correlated with potential outcome. Table 6

¹⁷ To complete the matriculation process, many requirements that tend to vary by school type and level of proficiency in each subject. The verification of information between the administration and the schools is a lengthy process. The first version of the matriculation data file becomes available in October and the final in December of the same year.

¹⁸ The sample was limited to schools with positive ($> 5\%$) matriculation rates.

presents the estimated coefficients from regression of the measurement error on student characteristics, lagged students' outcomes and school characteristics. These regressions were run with school level means of all variables, separately for the whole sample (507 high schools) and only for the eligible sample (97 schools). The whole set of regressions were estimated twice, once with data of the 2000 high school seniors and once with the data of the 2001 seniors.

The first panel of Table 6 presents twenty estimated coefficients from regressions of the 1999 measurement error on student's characteristics; only *one* of these estimates is significantly different from zero (the coefficient on percent of immigrant students in the sample of eligible schools in year 2001). The second panel in table presents twenty-four estimated coefficients from regressions of the 1999 measurement error on student's pre-program outcomes, only three of which are marginally significantly different from zero. Based on the evidence presented in Figure 1-4 and in Table 6, it may safely be concluded that the 1999 measurement error does not correlate with observable characteristics that may correlate with potential outcomes.¹⁹

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 the participation status, with $T = 1$ for participants and $T = 0$ for non-participants. 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 measurement error can be used for identification either as the basis for structuring a natural experiment, where treatment is assigned randomly in a subsample of the ninety-seven-school sample or as an instrumental variable. Seventeen of the forty-nine treated schools had a correct 1999 matriculation rate above the threshold line. Thus, these schools were "erroneously" chosen for the program. For each of them, there might have been a school with a similar matriculation rate but with a random measurement error not large (and negative) enough to drop it below the assignment threshold. This amounts to non-parametrically matching schools on the basis of the value of S^* . Figure 5 shows this pairing. The drawn ellipse circles the treated schools and their matching counterparts. There are twelve such ellipses. Within this sample (twenty-nine schools) treatment assignment was random, as shown above. Therefore, the twelve untreated schools may be used as a control group that reflects the

¹⁹ Another possible way to test for the random nature of the measurement error was to test if it is serially uncorrelated. However, lacking more than one year of data on the initial matriculation rate and its revised value, I so cannot compute the measurement error for any previous years.

counterfactual for identification of the effect of the program. The treated schools in this sample, however, are not a random sample culled from the sample of all treated schools, as may be seen clearly in Figure 5. For example, the correct 1999 matriculation rate is 45 percent or higher for all schools in this sample, while many schools in the full sample have correct matriculation rates that is lower than 45 percent. One should bear this in mind when interpreting the results, especially in the case of treatment heterogeneity. Importantly, however, the range of the 2000 matriculation rates in this sample is much wider (for both participating and non-participating schools); it ranges from 32 to 79 percent. This wider range of the school matriculation rate mitigates to some extent the limitation in terms of external validity of the findings that are based on the RD natural experiment sample.

Table 7 presents the pre-program (2000) and post-program (2001) means of students and school characteristics for the seventeen treated schools and the twelve control schools. The treatment–control differences and standard errors in these variables (columns 3 and 6) reveal that the two groups are very similar in both years in all background characteristics and in no case are statistically different. The only non-identical variable is number of siblings, and in 2001 the difference in the number of siblings was surprisingly large.

The second panel in Table 7 presents students' lagged outcomes for the 2001 senior students. These should be viewed as pre-program outcomes. No significant treatment–control differences are observed in English and math, in either year. Some differences are observed in history but not in biology or in total credits. The differences in history are evident in both years, but they probably reflect differences among schools in the timing of the history exam (in eleventh or twelfth grade), which is left to the discretion of the school.

The third panel in Table 7 compares the school-level covariates. Treatment and control are balanced in terms of religious status but not in terms of nationality, since there are no Arab schools in the control group. The 1999 mean matriculation rate is almost identical in the two groups, an unsurprising result since this school-level outcome was used for matching. A similar balance is found in the groups' 2000 matriculation rates.

The evidence in Table 7 suggests that, generally speaking, the treatment and control schools are well balanced in most student and school characteristics. Nevertheless, it is still necessary to control for all these variables in the estimation to net out the effect of any remaining differences. Furthermore, having similar data for the senior class in the year before treatment (2000) allows us to estimate a model with fixed school-level fixed school-level effects by using stacked panel data, which will absorb any remaining permanent differences, observed and unobserved, between the treated and the

control schools. The treatment effect estimated from this model is a difference-in-differences estimate embedded in a natural experiment setting.

Estimation and Results

Applying equation (1) to a school-level panel data structure with fixed school-level effects, the following model was used as the basis for regression estimates based on the RD natural experiment sample:

$$(3) \quad Y_{ijt} = \alpha + X_{ijt}'\beta + Z_{jt}'\gamma + \delta T_{ijt} + \Phi_j + \eta D_t + \varepsilon_{ijt}$$

where i indexes students; j indexes schools; t indexes years 2000 and 2001, and T is the assigned treatment status. As in equation (1), X and Z are vectors of student and school level covariates. This model also includes a constant effect for each year (D_t) with a factor loading η . The treatment indicator in this model is equal to the interaction between a dummy for treated schools and a dummy for year 2001 (T_{ijt} in equation [3] is equal to 1 for treated schools in year 2001 and 0 otherwise). The regressions were estimated using pooled data from both years (the two adjacent cohorts of year 2000 and 2001), stacked as school panel data with fixed school-level effects (Φ_j) included in the regression.

Table 8 presents the evidence for two different specification of equation 3, with and without the 1999 correct matriculation rate included as a control. The standard errors reported in the table are adjusted for clustering, using formulas set forth in Liang and Zeger (1986).²⁰

The treatment effect in English and math, for all three outcomes, is positive but varies in degree of precision. In math, all three outcomes are significantly different from zero and in English only the estimated treatment effect on attempted exams is not large enough relative to its estimated standard error. The effect of treatment on credits earned in math is 0.256, a 18 percent improvement relative to the mean of the control schools (1.46). The effect of treatment on awarded credits in English is 0.361, a 17 percent improvement relative to the mean of the control schools (2.11). The relative improvement in credits attempted is much lower—7.0 percent in math and 8.4 percent in English. These results reinforce the pattern observed in the OLS and PSM results reported above, namely that the effect of teachers' incentives works through two channels: the first increases the

²⁰ A disadvantage of the Liang and Zeger method is that the validity of Generalized Estimating Equation inference turns on an asymptotic argument based on the number of clusters. The sample of thirty schools may be considered too small for asymptotic formulas to provide accurate approximation to the finite-sample sampling distribution (Thornquist and Anderson; 1992). However, since I am using school panel data, the number of clusters is twice the number of schools since the unit of clustering is defined as the interaction of school and year. Therefore, the number of clusters is sixty.

attempt rate of exams and credits; the second increases the passing rate. The sizes of the two types of effects suggest that in the RD natural experiment sample the latter is the more important.

The interpretation of the foregoing results as causal is based on the random assignment of program status by the measurement error, conditional on the actual 1999-matriculation rate. Indeed, the treatment-effect estimates are sensitive to the exclusion of the correct 1999 matriculation rate as a control. Without this control, for example, the estimated effect of treatment on math credits earned is much lower, 0.163 versus 0.256, and is less precisely estimated. The English treatment estimate is also lower but only marginally.

The results presented above resemble in magnitude some of the estimated treatment effects obtained by the use of the OLS and PSM methods (reported in Tables 2 and 4, respectively). The effect of treatment on math credits earned, for example, is 0.293 in the PSM method and 0.256 in the RD natural experiment method. The estimates of the effect on English credits attempted are also almost identical, 0.224 and 0.230, respectively. However, some of the other estimates are different; e.g., the effect on English credits earned is 0.145 in the PSM method, smaller than the estimate in the RD natural experiment method (0.361).

I also used the RD natural experiment method to estimate models allowing treatment effects to vary with lagged outcomes. Using the quartile dummies described in the previous section, I estimated the following model for each of the outcomes:

$$(4) \quad Y_{ijt} = \alpha + X_{ijt}'\beta + Z_{jt}'\gamma + \sum_q d_{qi}\mu_q + \sum_q \delta_q T_{jt} + \Phi_j + \eta D_t + \varepsilon_{ijt},$$

where δ_q is a quartile-specific treatment effect and μ_q is a quartile main effect. Significant estimates are expected mainly for students in quartiles 1–3.

Table 9 reports results of the estimation of equation (4), using the RD natural experiment sample. Significant positive effects on number of exams and credits attempted are estimated, as expected, only for students in the first and second quartiles. The quartile pattern of the effect on the passing rate in the exams (credits awarded) also reveals a significant positive effect in the third quartile in math but not in English. The largest absolute effect on credits awarded is in the second quartile. The effect on math is an increase of a half a credit, against a mean of one credit in the control group, an impressive 50 percent increase. In English the effect in the second quartile is an increase of 0.58 credits against a mean of 2 credits in the control group, implying an increase of 30 percent due to the program. However, the most dramatic effects on credits awarded are in the first quartile: a 74 percent increase in math (a 0.258 change against the control group mean of 0.347) and a 78 percent increase (a 0.707 change against the control group mean of 0.911) in English.

The last panel in Table 9 presents the effect of the program, by quartile, on the matriculation rate. A positive and significant effect is estimated only for the second quartile, a 7.6 percent increase, which implies a 20 percent improvement against a 38.6 percent counterfactual, the mean of the second quartile of the control group. No effect on the matriculation rate is found in the first quartile.

4.2 A Sharp Regression Discontinuity Design

Since the rule governing selection to the program was based simply on a discontinuous function of a school observable (the erroneously measured 1999 matriculation rate), the probability of receiving treatment changes discontinuously as a function of this observable. This sharp discontinuity in the treatment assignment mechanism may be exploited as second RD identification information for evaluation of the effects of the teachers' bonus program.²¹ The discontinuity in our case is a sharp decrease (to zero) in the probability of treatment beyond a 45 percent school matriculation rate for nonreligious Jewish schools and beyond 43 percent for Jewish religious schools and Arab schools. 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 participants may be similar to marginal nonparticipants. (In this context, the term "marginal" refers to those schools that are not too far from the selection threshold.) The degree of similarity probably 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 is not likely to be related to random changes. Therefore, a bandwidth of about 10 percent seems to be a reasonable choice in our case.

This identification strategy may be presented as follows: Let r be the threshold for participation ($r=45$ or $r=43$), so that $I=1(S \leq r)$. The participation status for schools in a neighbourhood of r changes for non-behavioral reasons. Marginally participant (r^-) and marginally non-participant (r^+) schools define "quasi-experimental" groups. The main drawback of this approach is that it allows us to estimate the effect for marginally exposed schools only. In the presence of heterogeneous impacts, it allows us only to identify the mean impact of the intervention at the selection threshold, which may be different from the effect for schools that are far from the threshold for selection.

²¹ Regression discontinuity designs were described by Campbell (1969) and were formally examined as an identification strategy recently by Hahn, Todd, and van der Klaauw (2001). For recent examples, see Angrist and Lavy (1999, 2002a), Lavy (2002), and van der Klaauw (1997).

There are twelve untreated schools with matriculation rates in the 0.46–0.52 range and fourteen treated schools in the 0.40–0.45 range (Figure 6). 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 RD natural experiment sample. Nine of the fourteen treated schools and five of the twelve control schools belong to the groups of treatment and control schools, respectively, in the RD natural experiment sample. However, note that twelve of the twenty-six schools (almost 50 percent) included in the sharp RD sample were not among the thirty schools that make up the RD natural experiment sample. This suggests that the overlap between the two samples still leaves enough “informational value added” in each of the samples.

Table 10 replicates Table 7 for the sharp RD sample. The treatment–control differences and standard errors in the student’s background variables (columns 3 and 6) reveal that the two groups are very similar in both years in all characteristics except the ethnicity variable. The proportion of treated students of African-Asian origin is lower in treated schools than in control schools; this difference is significant in 2000 but not in 2001. The second panel reveals some control-treatment differences in the lagged attempt rate of exams and credits in some subjects but only a few of these estimates are marginally significant. The third panel reveals a statistically significant treatment–control gap in the erroneously measured 1999 matriculation rate and a similar gap in the correct rate. The gap carries the expected sign, negative, because all treated schools had erroneously measured matriculation rates below the threshold value and all control schools were above the threshold. Given that all measurement errors in the discontinuity sample were negative, the treatment–control difference in the correct matriculation rate is expected to be negative as well. The two differences are of similar magnitudes—0.061 and –0.054—and both had low standard errors. The estimation below will include as controls all the variables in Table 10. However, since the measured differences may reflect other unmeasured differences, identification based on the sharp RD approach depends more than in the case of the RD natural experiment approach on the school constant effects model to net out any remaining unobserved fixed correlates of potential outcomes. The lagged outcomes that are included as controls also increase the likelihood that many of the remaining confounding factors will be netted out.

Estimation Results

The sharp RD sample was used to estimate models identical to those estimated with the RD natural experiment sample (equations [3] and [4]). In principle, the identification based on the sharp RD 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 and therefore the latter should not be included as a control.

In contrast to the estimates based on the RD natural experiment sample, there is no reason to expect the results based on the sharp RD sample to be sensitive to control for the lagged true matriculation rate. However, for the purpose of comparison I again estimated two specifications, with and without controlling for the correct 1999 and 2000 matriculation rates even though identification is not conditioned on this variable.

Table 11 presents the results. The treatment-effect estimates are very similar though always lower than those obtained using the RD natural experiment sample. The estimates based on the sharp RD are expected to be downward-biased because the control group has higher average pre-program outcomes. The treatment-effect estimates are positive and significantly different from zero for all English and math outcomes except for the number of math attempted-exams outcome, which is only marginally significant. The estimated effect on earned credits in math is 0.244 (S.E.=0.078), just slightly lower in size and precision than the estimate obtained with the RD natural experiment sample. The estimated effect on English credits earned is 0.177 (S.E. = 0.104), about half the estimate derived from the RD natural experiment and less precisely estimated.

Note that the treatment estimates in Table 11 are not sensitive at all to the exclusion of the true 1999 and 2000 matriculation rate as a control variable; the coefficients in the second row of Table 11 are practically identical to those presented in the first row. This result suggests that the sensitivity of the results in the RD natural experiment sample (Table 8) to control of the lagged correct matriculation rate was unique to that sample. This result, which may be viewed as a specification check, strengthens the credibility of the causal interpretation of the RD natural experiment results, especially given the >50% overlap between the two samples.

Table 12 reports results of the estimation of equation (4), which allows treatment to vary by quartile of the distribution of the average score in pre-program matriculation exams, using the “discontinuity” sample. The pattern in the table is qualitatively very similar to that of Table 9: almost no significant effects are estimated for students with above-average lagged performance (quartiles 3 and 4) and the highest effects are estimated for the second quartile. The effects on the math outcomes are also quantitatively similar to those reported in Table 9. However, the results regarding the effect on the English outcomes are somewhat lower than those reported in Table 9. The effect on the matriculation rate of students is again significant for the second quartile, at an increase of 8.9 percent,

which is not very different from the 7.6 percent effect shown in Tables 9. However, small positive effect is also evident in the first quartile, similar to the respective OLS and MPS results.

5. Spillover Effects of the Program

Incentives may induce strategic behavior. For example, teachers may prompt students to reallocate their time and effort toward the rewarded subjects at the expense of other subjects. Hence, the program may have an adverse effect on outcomes in subjects other than those rewarded. However, additional effort on the part of teachers in the program may actually free up some of the students' time for other subjects. In such a case, the effect on outcomes in other subjects may actually be positive. This potential spillover or substitution aspect of the program may be addressed by estimating the effect of the program on the outcomes of all other "untreated" subjects. However, the number of subjects that may be considered truly untreated is limited, for two reasons. First, students are tested in many different subjects at the end of twelfth grade and the sample size in some of the tests is very small. Second, one should bear in mind that schools were allowed to include in the program teachers of one other subject in lieu of Hebrew or Arabic. Therefore, I confine the focus to untreated subjects that had the largest sample size—history and biology—and I estimate the effect of the program on the overall number of attempted exams, attempted credits, and earned credits in all untreated subjects.

Table 13 presents OLS, PSM, RD-natural experiment, and sharp RD estimates of the three outcomes in history and biology and the three outcomes in all untreated subjects. The estimates vary by the methods of estimation used. The OLS and the PSM estimates have a very similar pattern, suggesting that the incentive program induced students to take more exams and attempt to acquire more credits in history and biology as well as in other untreated subjects, but did not lead to a higher success rate in any of the untreated subjects. The estimated effects on attempted exams and attempted credits in both history and biology, for example, are positive and marginally significant with t values of 1.7–1.8. The size of these effects is relatively large, 25–30 percent of the control-group sample means of these outcomes. On the other hand, the point estimates on earned credits are practically zero in both subjects: the estimated treatment effect is 0.027 (S.E. = .046) on history credits earned and 0.054 (S.E.=0.019) on biology credits earned. The estimated treatment effect on attempted credits in all untreated subjects is 0.468 and its estimated standard error is 0.242. The estimated treatment effect on earned credits in all untreated subjects is 0.054 and its estimated standard error is 0.193. On the basis of this evidence, it seems that the program led to some increase in the number of attempted credits in untreated subjects but had no effect on the respective passing rate. These results do not

change when the models estimated allow for heterogeneity in the effect of treatment by student ability.

The evidence on spillover effects is less clear-cut even when estimated on the basis of both versions of the RD method. The estimated effects on the biology and history outcomes are very imprecisely estimated and vary in signs. On the other hand, the effect on the attempt rate of exams and credits in all untreated subjects together is positive though only marginally significant. When heterogeneity in the effect of treatment by student ability was allowed, significant effects on these outcomes were estimated for students in the second quartile but not for other students. In the second quartile, the estimated effect on total credits attempted in all untreated subjects is 0.429 and its standard error is 0.323; the respective estimates for total credits earned is 0.578 and 0.331. For all other three quartiles the estimated effects are practically zero. These results, as in the case of the OLS and PSM estimates, may be viewed as some evidence of spillover effects of the program for students in the second quartile of ability.

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

The evidence in the previous section shows clearly that the teachers' incentive program led to significant improvements in students' achievements in English and math. How closely do these improvements correspond to greater effort on teachers' part? Do they reflect different pedagogy and teaching methods? 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.²² To address these questions, a telephone survey was conducted among the English and math teachers who participated in the program.²³ For comparison purposes, a similar survey was conducted with a similar number of nonparticipating English and math teachers. Table A1 in the Appendix shows that the characteristics of the teachers in the two groups are very similar.

Table 14 presents evidence about the effect of the incentive program on three behavioral outcomes of participating teachers: teaching methods, teachers' effort, and focusing of effort on weak

²² See, for example, Glewwe, Ilias, and Kremer, 2003.

²³ 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.

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

The evidence, shown for English and math teachers separately, points to two patterns: the program modified teaching methods and led to a major increase in teachers' 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.

The proportion of program-participant English teachers who taught in small groups is 12.2 percentage points higher than the respective proportion (59.6 percent) among nonparticipants (Table 14). More dramatic and significant differences are evident with respect to the proportion of teachers who use individualized instruction and tracking in the classroom by ability: 71 percent and 75 percent among participating teachers, respectively, as against 56 percent and 43 percent, respectively, among comparison-group teachers. Ninety-three percent of comparison group teachers reported that they adapt teaching methods to their students' ability; 99 percent of treated teachers so reported. Among math teachers, the only significant difference in teaching methods is in the prevalence of tracking by ability; this practice was used by 56 percent of program teachers as against 40 percent of comparison-group teachers.

One-third of English teachers in the comparison group, as against 37.4 percent of participating teachers, reported that they added special instruction time throughout the school year, even before the program started. Among math teachers, about 50 percent of participating and nonparticipating teachers added instruction time beyond their regular teaching load. This evidence may be regarded as pre-program baseline evidence about teachers' effort because the program did not begin until January 2001. However, the answer to the question about added instruction time during the exam preparation period, from mid March to the end of June, reveals significant treatment-control differences. The difference is very large, at 20 percentage points (41.1 percent versus 21.1 percent) among English teachers, and 8 percentage points (37 percent versus 29.1 percent) among math teachers. These differences are significantly different from zero in both subjects. No significant

difference was found among English teachers in terms of the amount of instruction time added, about five hours a week in both the treatment and the control groups. Among math teachers, however, there was a significant treatment–control difference in this parameter, at 5 versus 6.7 hours weekly hours, almost a 35 percent difference. Table 14 however, reveals also that the targeting of effort to the weakest students is much more prevalent among participating English teachers (39 percent) than among nonparticipating teachers (31 percent). Participating math teachers, on the other hand, direct their effort more toward average and strong students.

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.

Further evidence that may be relevant, albeit indirectly, to the issue of teaching to the test is the impartiality of grading behavior of teachers in the program. The final grade in each of matriculation subject is an average of two scores: an internal school score given by the subject teacher and an external score based on a national exam. Comparison of these two scores shows that there are no significant differences between program and non-program teachers in terms of the gap between the two scores in English and math. This means that participating teachers did not inflate, in comparison to other teachers, the grades that they gave to their students relative to their ability as reflected in the external score. Since I showed in Sections 4–6 that students in the program outperformed nonparticipating students on the external exams, this implies that the higher absolute scores that participating teachers gave their students were matched by an equally improved performance on the external exams.

7. Does Tournament Ranking Correlate with Teachers’ Characteristics?

The results presented in this paper indicate that individual teachers matter in improving schooling quality. 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. Table A1 presents such correlations for math and English teachers.

The results support the view that teaching quality is not highly correlated with characteristics such as age, gender, education, teaching certification, and years of teaching experience.²⁴ None of these variables was very significant in explaining the ranking of teachers in the tournament.²⁵ Other variables, however, showed significant correlations in the 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 (the 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.

8. A Cost-Benefit Comparison with Alternative Interventions

The teacher-incentive program cost \$170 per student and improved the matriculation rate by about 4 percentage points. The student-bonus program evaluated by Angrist and Lavy (2002) cost \$300 per student and elevated the matriculation rate by 7 percentage points. The school-incentive program analyzed by Lavy (2002) cost \$270 per student and boosted the matriculation rate by 1–2 percentage points. Among the three incentive programs, the student-bonus program was the most expensive in per-student terms but it was as effective in cost-equivalence terms as the teacher-bonus program when adjusted for its higher impact on the matriculation rate. The group-school incentive program was the least effective in cost-equivalence terms among the three incentive-based programs. All these programs were more effective than a program that targeted additional instruction time to underachieving students in small groups (two to six students) in several matriculation subjects.²⁶

Another way of benchmarking costs and benefits is by comparing the program cost per student to the likely economic benefits of the improved outcomes, e.g., of earning a matriculation certificate, with the cost of about \$170 per treated student that led to a 4 percentage-point increase in the matriculation rate. One may measure the economic benefit of having a matriculation certificate. Angrist and Lavy (2002) estimated the economic benefit of a matriculation certificate to an individual

²⁴ See Heckman (2002) and Hanushek (2002) for discussion of this point.

²⁵ None of the teachers' attributes was significantly correlated with any of measure of teachers' effort discussed in the previous section.

²⁶ The cost of this program was \$1,100 per student and it raised the matriculation rate by 11 percentage points (Ministry of Education, Evaluation Division, May 2002).

who has twelve years of schooling at \$4,025 per year. Given that the teacher-bonus experiment raised the mean probability of matriculation among treated students by 4 percent, it should increase the annual earnings of treated students by $4,025 \times .04 = \$161$ per person per year. This allows the program to recoup its cost quickly, just after two years.

Finally, it is worth noting that teachers' incentives, beyond affecting motivation, may also have a long-term effect by means of the screening and selection of teachers (Lazear, 2000 and 2001). Pay for performance will result in higher pay for the better teachers, which may encourage the right pattern of retention and turnover of teachers through selection. In other words, a pay-for-performance scheme may lead to a different and more productive applicant pool from which teachers are selected. Estimating such a long-term effect is not feasible in this study.

9. Conclusions

The evidence presented in this paper indicates clearly that pay-for-performance incentives “work” as well among schoolteachers as they do among practitioners of other occupations. 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 output is 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 teachers' incentives are a very promising path toward the improvement of school quality. The results of this new experiment add important evidence to those of Lavy (2002) concerning group school incentives.

The non-random nature of the assignment of schools and teachers to the experiment entailed alternative identification strategies. The two variants of the regression discontinuity that I used for identification—the natural experiment, based on a random measurement error in the assignment variable, and the sharp regression discontinuity—provided appealing approaches to the problem of non-random assignment into the incentive pay program. The propensity matching method that I also used seems very appropriate in our case because of the rich data that may be used for matching, especially test scores from many pre-program high school exams, which are very likely to correlate with unobserved student attributes such as ability and motivation. All the methods of identification used in this study yielded qualitatively similar results.

Incentive programs in schools may include all teachers or only those who teach specified subjects. Targeting teachers of some subjects may be appropriate if the objective is, for example, to enhance the outcome in some subjects or to cope with budget constraints. The targeting of incentives

to some teachers only, however, may have the drawbacks of jeopardizing the solidarity of the teachers in a school and generating negative spillover and externalities. The potential positive effect of incentives on targeted outcomes may have unintended negative (or positive) effects on other outcomes that may outweigh (or enhance) the positive effect of the incentives on intended outcomes. In this experiment, I had a unique opportunity to examine the existence and size of spillover effects because the program aimed the financial incentives at only some schoolteachers. The evidence presented in this paper suggests that this targeting of incentives led to small and marginally significant spillover effects. The spillover effects were more pronounced when their impact was allowed to vary by student's ability but they were positive and significant only for students in the two lowest ability quartiles. However, the caveat of this result, and for that matter also of the other results presented in this paper, is that the experiment lasted for just one year and, therefore, did not allow us to study the long-term effects of the incentives.

On the basis of a post-program survey among participating and nonparticipating teachers, I found evidence that links the improvement in students' cognitive outcomes on the 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. This resemblance makes the results and lessons of the incentive experiment examined in this paper relevant for many education systems in Europe and the US.

10. References

- Angrist, J. (1998). "Using Social Security Data on Military Applicants to Estimate the Effect of Military Service Earnings." *Econometrica* 66 (2): 249-288.
- Angrist, J. and Lavy, V. (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics*. 114 (2): 533-575.
- Angrist, J. and Lavy, V. (2002a). "New Evidence on Computers in the Classroom." *The Economic Journal*, October 2002.
- Angrist, J. and Lavy, V. (2002b). "The Effect of High School Matriculation Awards: Evidence from Randomized Trials." NBER Working Paper Number 9389.
- Campbell, D. T., "Reforms as Experiments," *American Psychologist* 24 (1969), 409-429.
- Cohen, D. and R. Murnane (1985). "The Merits of Merit Pay." *Public Interest*, 80 summer: 3-30.
- Clotfeller, C. T., and H. F. Ladd. (1996). "Recognition and Rewarding Success in Public Schools." In: H. F. Ladd (ed.), *Holding Schools Accountable: Performance-Based Reform in Education*. Washington, D.C.: Brookings Institution.
- Dearden, L., C. Emmerson, C. Frayne and C. Meghir (2003). "The Impact of Financial Incentives On Education Choice," Presented in a CEPR conference "The Economics of Education and Inequality," May 2003.
- Elmore R. F, C. H. Abelman and S. H. Fuhrman (1996). "The New Accountability in State Education Reform: From Process to Performance." In: H. F. Ladd (ed.), *Holding Schools Accountable: Performance-Based Reform in Education*. Washington D.C.: Brookings Institution.
- Gaynor, Martin, and Mark V. Pauly (1990). "Compensation and Productive Efficiency in Partnership: Evidence from Medical Group Practice." *Journal of Political Economy* 98(3): 544-73.
- Gibbons, Robert (1998). "Incentives in Organizations," *Journal of Economic Perspectives* 12(4): 115-132.
- Glewwe, Paul, N. Ilias and M. Kremer (2003). "Teacher Incentives," NBER Working Paper Number W9671.
- Green, Jerry and Nancy L. Stokey (1983). "A Comparison of Tournaments and Contracts." *Journal of Political Economy* 91: 349-64
- Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. "Identification and Estimation of treatment Effects with a Regression-discontinuity Design." *Econometrica* 69 (2001):201-209.
- Hanushek, E. (2002). "Publicly Provided Education," NBER Working Paper No. 8799.
- Hards, E. C. and T. M. Sheu (1992). "The South Carolina School Incentive Reward Program: A Policy Analysis." *Economics of Education Review*, Vol. 11, No. 1: 71-86.
- Heckman, J. J., (2002). "Human capital Policy," draft.
- Heckman, J. J., H. Ichimura and P. E. Todd (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, 64 (4): 605-54.
- Heckman, J. J., H. Ichimura and P. E. Todd (1998). "Matching as an Econometric Evaluation Estimator: Evidence," *Review of Economic Studies*, 65 (2): 261-294.

- Holmstrom, B. and P. Milgrom (1991). "Multitask Principal-Agent Analysis: Incentive Contracts, Asset Ownership and Job Design," *Journal of Law, Economics and Organization* 7 (Special Issue), 24–52.
- Israel Ministry of Education, Bagrut Test Data 2000, Jerusalem: Ministry of Education, Chief Scientist's Office, April 2001.
- Israel Ministry of Education, "The Bagrut 2001 program, an Evaluation". Jerusalem: Ministry of Education, Evaluation Division, May 2002.
- Jensen C. Michael and Kevin J. Murphy (1990). "Performance Pay and Top-Management Incentives," *The Journal of Political Economy* 98(2): 225-264.
- Kandel, E. and E. Lazear (1992). "Peer Pressure and Partnership." *Journal of Political Economy* 100 (4):801-17.
- Kelley, Carolyn and Protsik, Jean. (1996). Risk and Reward: Perspectives on the Implementation of Kentucky's School-Based Performance Award Program. American Educational Research Association conference paper, April 8, 1996, New York City.
- Lavy, V. (2002). "Evaluating the Effect of Teachers' Group Performance Incentives on Students Achievements." *Journal of Political Economy*, 10 (6), December 2002, 1286–1318.
- Lazear, E. and S. Rosen. (1981). "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy* 89: 841–64.
- Lazear, E. "Performance Pay and Productivity," *American Economic Review*, December, 2000.
- Lazear, E. "Paying Teachers for Performance: Incentives and Selection," Draft, August 2001.
- Liang, Kung-ye, and Scott L. Zeger, "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika* 73 (1986), 13–22.
- Malcomson, J. (1998): "Incentives Contracts in Labor Markets," in Ashenfelter, O. and D. Card, eds., *Handbook of Labor Economics* 3(B): 2291–2372.
- Milgrom, P. and J. Roberts (1992). *Economics, Organization and Management*, Prentice Hall, New Jersey.
- Moulton, B. "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics* 32 (1986), pp. 385–97.
- Prendergast, Canice. (1999). "The Provision of Incentives in Firms." *Journal of Economic Literature* 37: 7–63.
- Rosenbaum, P.R. and Rubin, D.B., (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.
- Rosenbaum, P.R. and Rubin, D.B., (1985), "Constructing a Comparison Group using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician* 39: 33–38.
- Thornquist, Mark D., and G.L. Anderson, "Small-Sample Properties of Generalized Estimating Equations in Group-Randomized Designs with Gaussian Response," Fred Hutchinson Cancer Research Center, Technical Report, 1992.
- Wakelyn, David J. (1996). The Politics of Compensation Reform: A Colorado Case Study. American Educational Finance Association conference paper, March 23, 1996.
- Van der Klaauw, W. (1996). "A Regression-Discontinuity Evaluation of the Effect of Financial Aid Offers on Enrollment," unpublished manuscript, New York University.

Figure 1: The Relationship Between the Correct and the Erroneously Measured 1999 Matriculation Rate
Sample=507 Schools

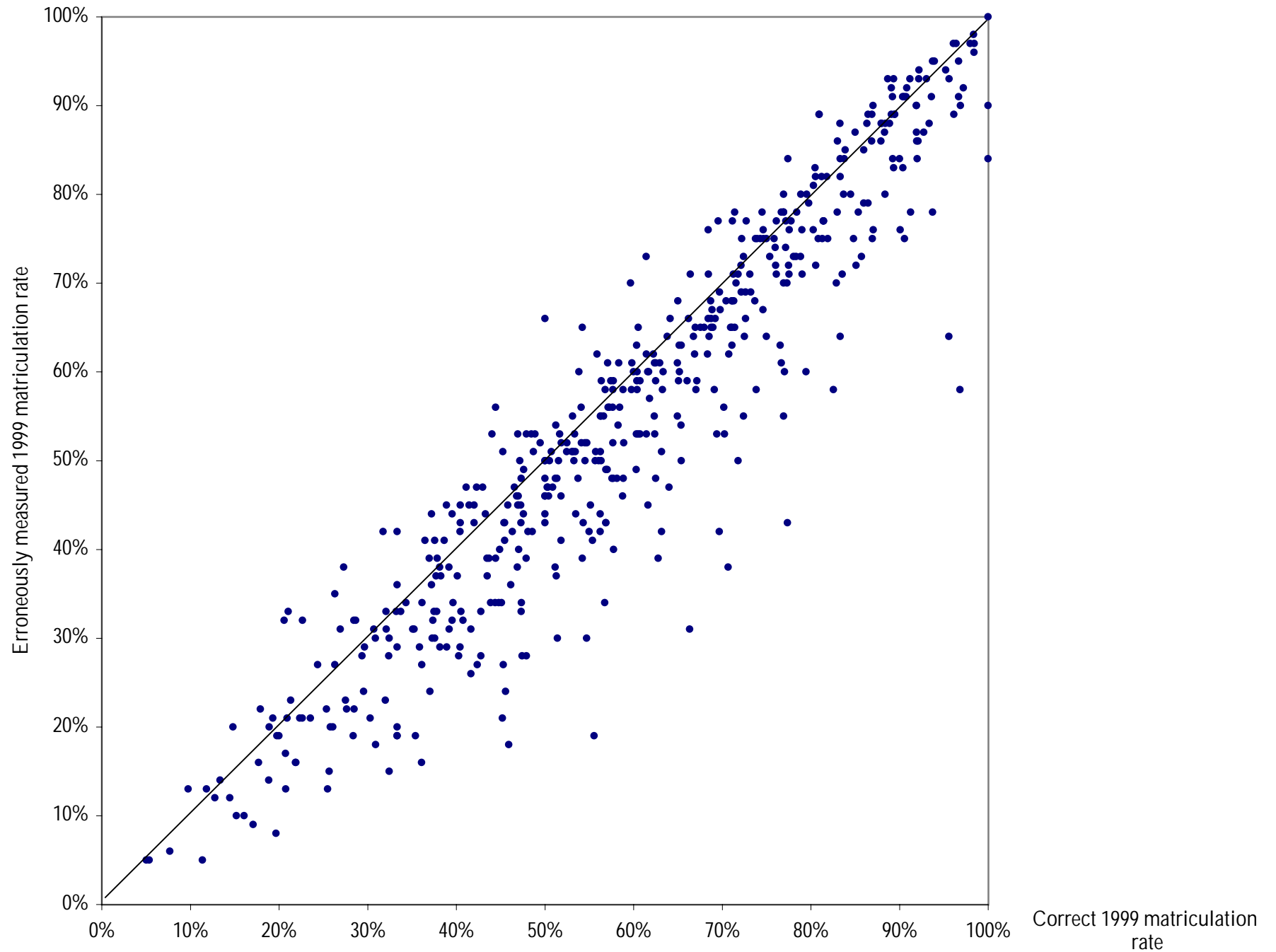


Figure 2: The Correct 1999 Matriculation Rate Versus The Measurement Error
Sample=507 Schools

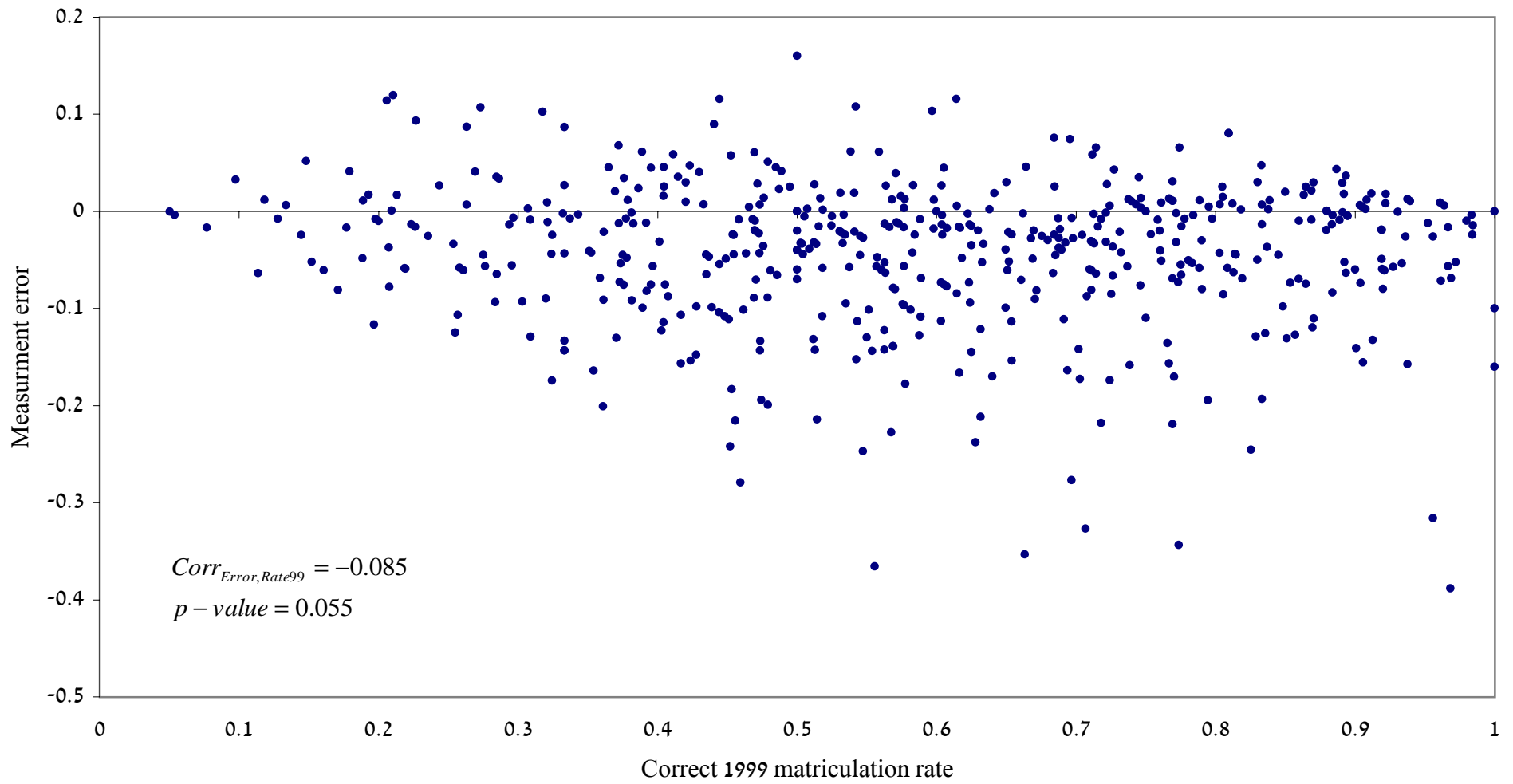


Figure 3: The Relationship Between the Correct and the Erroneously Measured 1999 Matriculation Rate
Sample=97 Schools

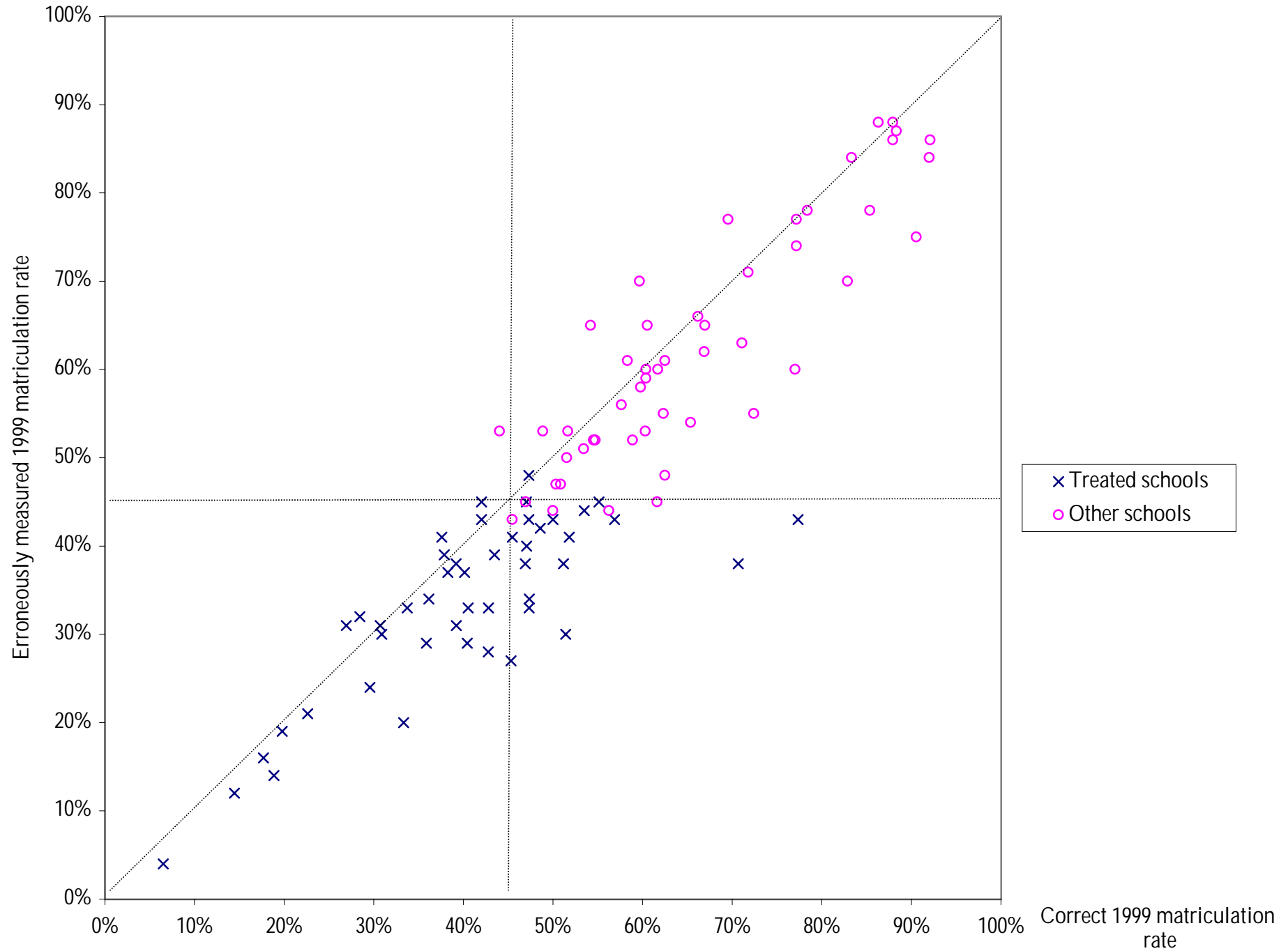


Figure 4: The Correct 1999 Matriculation Rate Versus The Measurement Error
Sample=97 Schools

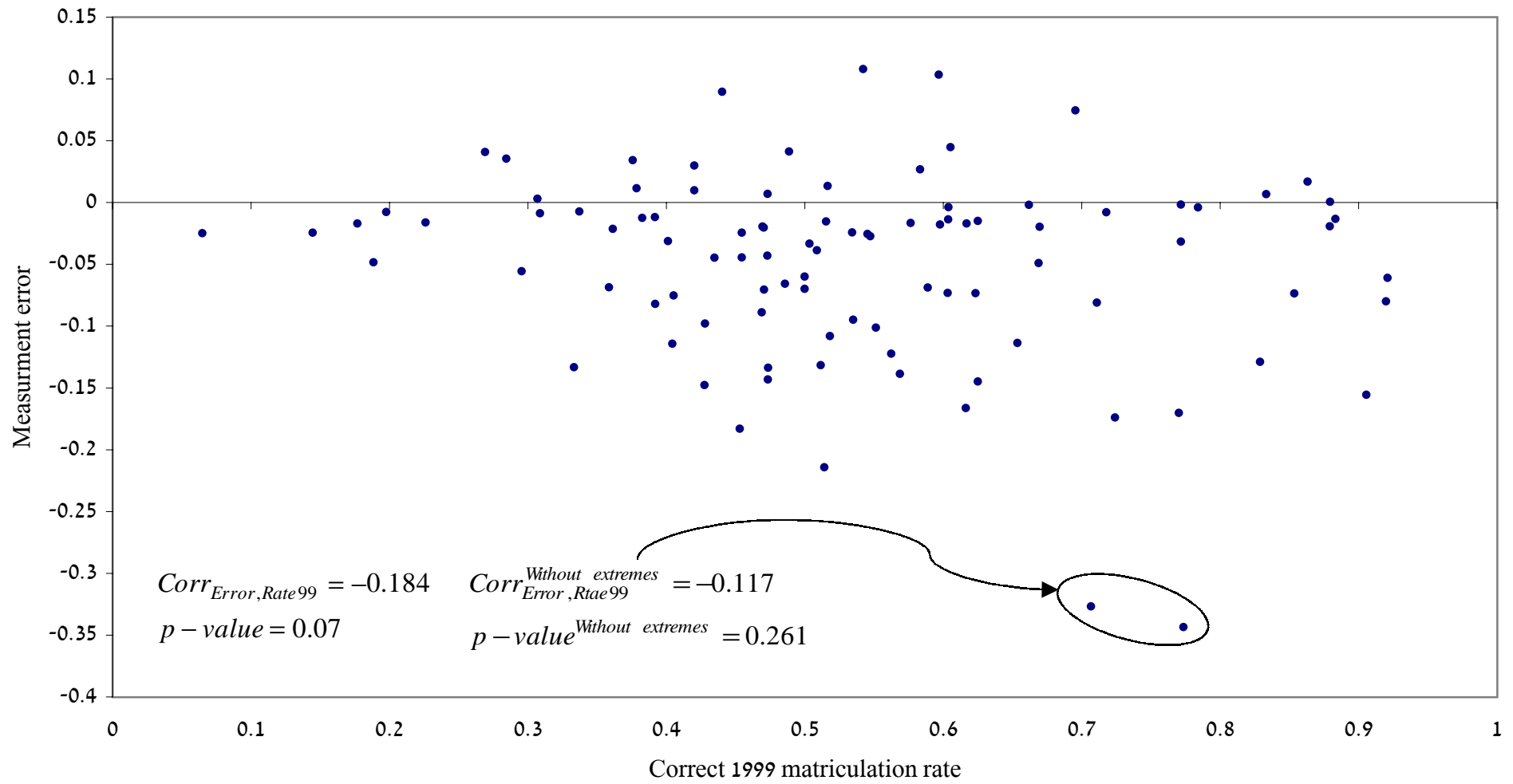


Figure 4A: The Correct 1999 Matriculation Rate Versus The Measurement Error
Sample=69 Schools

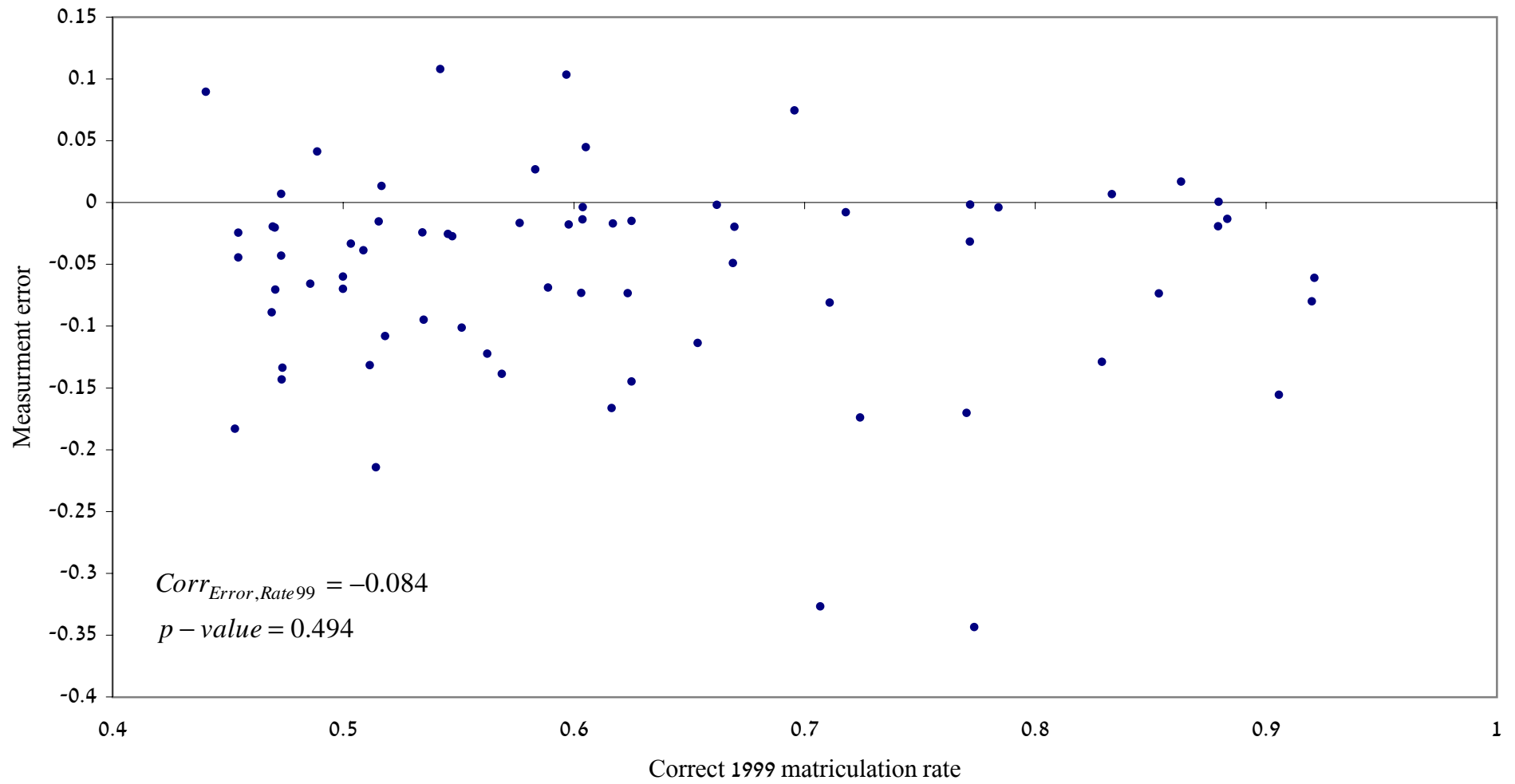


Figure 5: Determining the Sample of Schools That Were Randomly Assigned To Treatment or Control
Sample=97 Schools

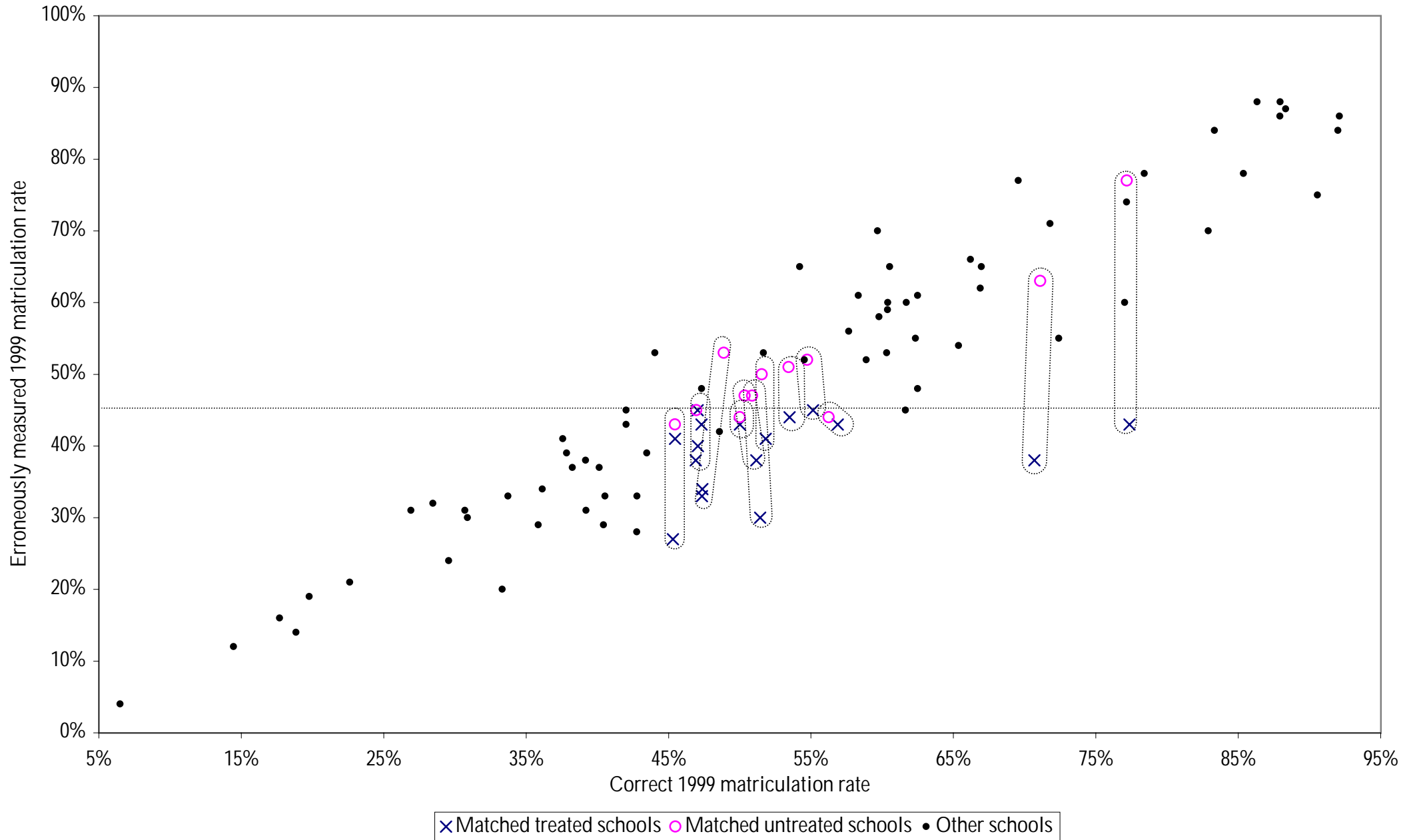


Figure 6: Determining the Discontinuity Sample (Schools Close To the Threshold Value)
Sample=97 Schools



Table 1

Descriptive Statistics: Program Schools Versus All Other High Schools

	Year 2000			Year 2001		
	All Schools	Program Schools	Difference (s.e)	All Schools	Program Schools	Difference (s.e)
Student's background						
Father's education	11.499	9.168	-2.330* (0.522)	11.097	9.160	-1.937* (0.487)
Mother's education	11.441	8.933	-2.508* (0.599)	10.853	8.701	-2.152* (0.618)
Number of siblings	2.666	3.421	0.755* (0.367)	2.618	3.403	0.786* (0.375)
Gender (Male=1)	0.467	0.495	0.028 (0.018)	0.466	0.483	0.017 (0.018)
Immigrant	0.024	0.030	0.006 (0.011)	0.015	0.022	0.006 (0.009)
Asia-Africa ethnicity	0.179	0.199	0.021 (0.024)	0.169	0.168	-0.001 (0.022)
Lagged student's outcomes						
Math credits	0.708	0.394	-0.314* (0.093)	0.774	0.427	-0.347 (0.092)
English credits	0.224	0.129	-0.095* (0.030)	0.228	0.112	-0.116 (0.037)
History credits	0.317	0.213	-0.104* (0.038)	0.594	0.362	-0.232 (0.056)
Biology credits	0.004	0.004	0.000 (0.004)	0.005	0.000	-0.005* (0.002)
Total credits	4.843	3.551	-1.291* (0.219)	5.018	3.588	-1.430* (0.230)
Average Score	69.362	57.227	-12.135* (1.800)	70.177	55.612	-14.565* (1.872)
School characteristics						
Religious school	0.167	0.195	0.028 (0.060)	0.168	0.180	0.012 (0.057)
Arab school	0.154	0.251	0.097 (0.070)	0.153	0.273	0.121 (0.074)
Previous year Bagrut rate	0.620	0.407	-0.213* (0.019)	0.625	0.410	-0.215* (0.021)
Number of observations	64,961	6,562		64,753	6,922	

Note: The table reports the mean of all variables by treatment and control, the differences of means and their standard errors adjusted for clustering using formulas in Liang and Zeger (1986). A * denotes significance level at 5% or 10%.

Table 2

The Treatment Effect Estimated By OLS

	Math			English		
	Control Group Mean	Program Effect Estimate	Standard Error Estimate	Control Group Mean	Program Effect Estimate	Standard Error Estimate
Attempted Exams	1.14	0.222	0.031	0.85	0.102	0.032
Attempted Credits	1.99	0.333	0.051	2.66	0.250	0.067
Awarded Credits	1.67	0.228	0.064	2.23	0.219	0.071
Bagrut Rate	0.62	0.054	0.018	0.62	0.042	0.017

Note: The table reports treatment-control differences for the three outcomes in English and Math. The sample used to produce the results reported in this table include the students that participated in the program and all the other students in the country: in math 72,051 and in English 72,378. Standard errors are adjusted for clustering at the school level using formulas in Liang and Zeger (1986) and presented in parenthesis. Students level controls include the number of siblings, gender dummy, father's and mother's education, a dummy indicating an immigrant student, a set of dummy variables for ethnic background, a set of dummies for the number of credit units gained in the relevant subject before treatment, overall credit units gained before treatment and the average score in the relevant tests.

Table 3

Descriptive Statistics: The Propensity Score Sample Based On Program Participants and their Matches

	Math matching			English matching		
	treatment	control	Difference (s.e)	treatment	control	Difference (s.e)
Student's background						
Father's education	9.282	9.297	-0.015 (0.550)	9.471	9.296	0.175 (0.541)
Mother's education	8.831	8.733	0.097 (0.697)	9.149	8.918	0.231 (0.673)
Number of siblings	3.342	3.483	-0.140 (0.394)	3.217	3.296	-0.079 (0.387)
Gender (Male=1)	0.488	0.471	0.017 (0.025)	0.483	0.471	0.012 (0.024)
Immigrant	0.014	0.001	0.014 (0.004)	0.014	0.001	0.014 (0.004)
Asia-Africa ethnicity	0.165	0.185	-0.021 (0.028)	0.179	0.210	-0.031 (0.027)
Lagged student's outcomes						
Math credits	0.362	0.304	0.058 (0.081)	0.531	0.535	-0.004 (0.109)
English credits	0.109	0.160	-0.051 (0.047)	0.060	0.051	0.009 (0.023)
History credits	0.442	0.409	0.033 (0.069)	0.453	0.419	0.034 (0.069)
Biology credits	0.000	0.002	-0.002 (0.001)	0.000	0.002	-0.002 (0.001)
Total credits	4.158	4.101	0.057 (0.288)	4.311	4.256	0.055 (0.259)
Average Score	64.225	63.416	0.809 (2.045)	65.047	63.559	1.488 (1.857)
School characteristics						
Religious school	0.179	0.203	-0.024 (0.064)	0.171	0.185	-0.015 (0.062)
Arab school	0.286	0.326	-0.041 (0.090)	0.232	0.253	-0.020 (0.081)
Previous year Bagrut rate (2000)	0.409	0.423	-0.014 (0.024)	0.422	0.426	-0.004 (0.023)
Number of observations	4,490	4,490		4,865	4,865	

Note: The table reports the mean of all variables by treatment and control, the differences of means and their standard errors adjusted for clustering using formulas in Liang and Zeger (1986).

Table 4

The Treatment Effect Estimated By the Propensity Score Matched Sample

	Math			English		
	Control Group Mean	Program Effect Estimate	Standard Error Estimate	Control Group Mean	Program Effect Estimate	Standard Error Estimate
Attempted Exams	1.13	0.242	0.032	1.00	0.088	0.033
Attempted Credits	1.98	0.398	0.059	2.54	0.230	0.068
Awarded Credits	1.58	0.293	0.068	2.06	0.145	0.079
Bagrut Rate	0.46	0.051	0.020	0.46	0.037	0.018

Note: The table reports treatment-control differences for the three outcomes in English and Math. The sample used to produce the results reported in this table are those of Table 3. Standard errors are adjusted for clustering at the school level using formulas in Liang and Zeger (1986) and presented in parenthesis. Students level controls include the number of siblings, gender dummy, father's and mother's education, a dummy indicating an immigrant student, a set of dummy variables for ethnic background, a set of dummies for the number of credit units gained in the relevant subject before treatment, overall credit units gained before treatment and the average score in the relevant tests.

Table 5

Effects on English and Math Bagrut Outcomes by Quartiles of Previous Test Scores, Propensity Score Matched Sample

	Estimates by quartile: Math				Estimates by quartile: English			
	1st quartile	2nd quartile	3rd quartile	4th quartile	1st quartile	2nd quartile	3rd quartile	4th quartile
Attempted exams								
Treatment effect	0.381 (0.063)	0.257 (0.046)	0.149 (0.039)	0.152 (0.051)	0.279 (0.058)	0.022 (0.049)	-0.004 (0.045)	0.046 (0.053)
Control group mean	0.582	1.179	1.341	1.431	0.851	1.194	1.054	0.915
Attempted credits								
Treatment effect	0.599 (0.099)	0.454 (0.075)	0.280 (0.079)	0.203 (0.105)	0.493 (0.105)	0.173 (0.087)	0.071 (0.092)	0.147 (0.133)
Control group mean	0.888	1.915	2.343	2.810	1.579	2.692	2.907	2.991
Awarded credits								
Treatment effect	0.341 (0.082)	0.396 (0.096)	0.274 (0.091)	0.106 (0.108)	0.245 (0.103)	0.175 (0.100)	0.000 (0.108)	0.119 (0.140)
Control group mean	0.452	1.353	1.927	2.600	0.995	2.097	2.488	2.673
Bagrut rate								
Treatment effect	0.041 (0.020)	0.104 (0.030)	0.026 (0.028)	0.047 (0.034)	0.031 (0.020)	0.086 (0.025)	0.018 (0.024)	0.034 (0.033)
Control group mean	0.060	0.373	0.654	0.762	0.060	0.382	0.658	0.766

Note: The table reports treatment effects for Math and English outcomes. The samples used in this table are identical to those used in Table 2. Treatment effects vary by quartile of summary Bagrut score through December 2000. Standard errors in parenthesis are adjusted for clustering at the school level using formulas in Liang and Zeger (1986). All models control for the student's and school characteristics that appear in Table 6 and also school fixed effects.

Table 6

Estimated Correlations Between the 1999 Measurement Error in the School Matriculation Rate
and Student's and School's Characteristics

	Year 2000		Year 2001	
	All Schools	Eligible Schools	All Schools	Eligible Schools
Student's background				
Father's education	0.001 (0.001)	-0.001 (0.002)	0.000 (0.001)	0.000 (0.003)
Mother's education	0.000 (0.001)	-0.001 (0.002)	-0.001 (0.001)	-0.003 (0.002)
Number of siblings	0.003 (0.002)	0.005 (0.004)	0.001 (0.002)	0.004 (0.004)
Gender (Male=1)	-0.003 (0.013)	-0.021 (0.033)	-0.005 (0.013)	-0.016 (0.033)
Immigrant	-0.035 (0.036)	-0.132 (0.093)	-0.002 (0.045)	-0.459* (0.164)
Lagged student's outcomes				
Math credits	-0.002 (0.004)	-0.011 (0.013)	0.000 (0.004)	0.001 (0.013)
English credits	0.006 (0.009)	0.066 (0.037)	0.013 (0.008)	0.089* (0.040)
History credits	0.004 (0.009)	0.050 (0.027)	0.004 (0.007)	0.028 (0.017)
Biology credits	-0.085 (0.059)	0.204 (0.204)	-0.116* (0.064)	0.148 (0.152)
Total credits	0.000 (0.002)	-0.004 (0.004)	0.001 (0.002)	0.002 (0.004)
Average score	0.000 (0.000)	0.000 (0.001)	0.0006* (0.0003)	0.001 (0.001)
School characteristics				
Religious schools	-0.006 (0.007)	-0.025 (0.016)	-0.006 (0.007)	-0.025 (0.016)
Arab school	0.025* (0.009)	0.024 (0.019)	0.025* (0.009)	0.024 (0.019)
Number of schools	507	97	507	97

Note: The coefficients presented in the table are based on regressions of the 1999 measurement error on student's characteristics and lagged Bagrut outcomes and school's characteristics. The data used are school sample means and regular standard errors are presented in parenthesis. A * denotes significance level at 5% or 10%.

Table 7

Descriptive Statistics: The Regression Discontinuity - Natural Experiment Sample

	Year 2000			Year 2001		
	Treatment	Control	Difference (s.e)	Treatment	Control	Difference (s.e)
Student's background						
Father's education	10.337	10.129	0.208 (1.007)	10.188	11.054	-0.865 (0.757)
Mother's education	10.315	10.340	-0.024 (1.061)	10.181	10.280	-0.099 (1.082)
Number of siblings	3.058	2.406	0.653* (0.351)	3.053	1.993	1.061* (0.389)
Gender (Male=1)	0.494	0.505	-0.011 (0.058)	0.534	0.517	0.018 (0.057)
Immigrant	0.017	0.029	-0.011 (0.027)	0.015	0.014	0.001 (0.014)
Asia-Africa ethnicity	0.228	0.287	-0.059 (0.057)	0.216	0.260	-0.045 (0.047)
Lagged student's outcomes						
Math credits	0.375	0.557	-0.182 (0.178)	0.320	0.583	-0.264 (0.153)
English credits	0.175	0.148	0.026 (0.060)	0.138	0.123	0.015 (0.083)
History credits	0.131	0.403	-0.271* (0.084)	0.353	0.775	-0.422* (0.161)
Biology credits	0.000	0.000	0.000 (0.000)	0.000	0.000	0.000 (0.000)
Total credits	4.055	4.256	-0.201 (0.443)	4.111	4.420	-0.309 (0.413)
School characteristics						
Religious school	0.296	0.205	0.091 (0.158)	0.307	0.206	0.102 (0.160)
Arab school	0.165	0.000	0.165 (0.098)	0.164	0.000	0.164 (0.100)
Previous year Bagrut rate (1999,2000)	0.501	0.552	-0.051 (0.032)	0.479	0.498	-0.019 (0.041)
Number of observations	2,405	1,773		2,350	1,678	

Note: The table reports the mean of all variables by treatment and control, the differences of means and their standard errors adjusted for clustering using formulas in Liang and Zeger (1986). A * denotes significance level at 5% or 10%.

Table 8

The Treatment Effect on English and Math Bagrut Outcomes Estimated Using the Regression Discontinuity - Natural Experiment Sample

	Math			English		
	Attempted exams	Attempted credits	Awarded credits	Attempted exams	Attempted credits	Awarded credits
Control group mean	1.04	1.93	1.46	0.94	2.66	2.11
Control for correct matriculation rate	0.078 (0.030)	0.135 (0.058)	0.256 (0.076)	0.027 (0.038)	0.224 (0.103)	0.361 (0.111)
No control for correct matriculation rate	0.079 (0.032)	0.125 (0.051)	0.163 (0.068)	0.035 (0.031)	0.194 (0.087)	0.327 (0.104)

Note: The table reports treatment-control differences for the three outcomes in English and Math. Standard errors are adjusted for clustering at the school level using formulas in Liang and Zeger (1986) and presented in parenthesis. The sample include 4,178 students of year 2000 and 4,028 students of year 2001 (same samples of Table 7). Students level controls include the number of siblings, gender dummy, father's and mother's education, a dummy indicating an immigrant student, a set of dummy variables for ethnic background, a set of dummies for the number of credit units gained in the relevant subject before treatment, overall credit units gained before treatment and the average score in the relevant tests. School fixed effects are included as well in each of the regressions.

Table 9

Effects on English and Math Bagrut Outcomes by Quartiles of Previous Test Scores, Using the Regression Discontinuity - Natural Experiment Sample

	Estimates by quartile: Math				Estimates by quartile: English			
	1st quartile	2nd quartile	3rd quartile	4th quartile	1st quartile	2nd quartile	3rd quartile	4th quartile
Attempted exams								
Treatment effect	0.081 (0.083)	0.184 (0.056)	0.017 (0.060)	-0.036 (0.067)	0.093 (0.095)	-0.009 (0.058)	-0.053 (0.073)	-0.009 (0.078)
Control group mean	0.506	0.949	1.186	1.373	0.733	1.114	1.024	0.858
Attempted credits								
Treatment effect	0.214 (0.123)	0.365 (0.105)	0.063 (0.097)	-0.141 (0.123)	0.497 (0.186)	0.357 (0.141)	0.020 (0.160)	-0.127 (0.205)
Control group mean	0.811	1.599	2.196	2.810	1.589	2.727	3.031	3.073
Awarded credits								
Treatment effect	0.258 (0.114)	0.499 (0.103)	0.334 (0.102)	-0.011 (0.114)	0.707 (0.160)	0.581 (0.151)	0.095 (0.153)	-0.084 (0.171)
Control group mean	0.347	0.973	1.627	2.550	0.911	2.007	2.500	2.746
Bagrut rate								
Treatment effect	0.027 (0.028)	0.076 (0.038)	0.013 (0.031)	-0.064 (0.035)	*	*	*	*
Control group mean	0.053	0.386	0.715	0.902	*	*	*	*

Note: The table reports treatment effects for Math and English outcomes. Treatment effects vary by quartile of summary Bagrut score through December 2000. Standard errors in parenthesis are adjusted for clustering at the school level using formulas in Liang and Zeger (1986). Sample identical to that of Table 8. All models control for student's and school characteristics, lagged outcomes and also school fixed effects.

* The estimates for the Bagrut rate are the same for Math and English because it is the same outcome in an identical sample of students.

Table 10

Descriptive Statistics: The Sharp Regression Discontinuity Sample

	Year 2000			Year 2001		
	Treatment	Control	Difference (s.e)	Treatment	Control	Difference (s.e)
Student's background						
Father's education	11.055	10.424	0.631 (0.486)	10.889	10.455	0.434 (0.511)
Mother's education	11.124	10.733	0.391 (0.561)	11.088	10.882	0.206 (0.575)
Number of siblings	2.609	2.427	0.182 (0.344)	2.552	2.131	0.421 (0.393)
Gender (Male=1)	0.492	0.468	0.024 (0.066)	0.498	0.488	0.010 (0.062)
Immigrant	0.013	0.045	-0.032 (0.022)	0.013	0.008	0.005 (0.007)
Asia-Africa ethnicity	0.218	0.319	-0.101 (0.050)	0.211	0.287	-0.077 (0.054)
Lagged student's outcomes						
Math credits	0.229	0.568	-0.339* (0.147)	0.265	0.578	-0.312 (0.176)
English credits	0.208	0.088	0.120* (0.061)	0.185	0.125	0.060 (0.090)
History credits	0.155	0.176	-0.022 (0.084)	0.434	0.567	-0.133 (0.149)
Biology credits	0.000	0.000	0.000 (0.000)	0.000	0.000	0.000 (0.000)
Total credits	4.044	4.499	-0.455 (0.346)	4.230	4.594	-0.364 (0.388)
School characteristics						
Religious school	0.098	0.325	-0.227 (0.152)	0.092	0.309	-0.217 (0.150)
Arab school	0.128	0.000	0.128 (0.090)	0.128	0.000	0.128 (0.091)
Previous year Bagrut rate (1999,2000)	0.483	0.537	-0.054 (0.016)	0.482	0.507	-0.025 (0.042)
1999 measured with error Bagrut rate	0.426	0.488	-0.061 (0.011)	-	-	-
Number of observations	2,523	1,564		2,535	1,406	

Note: The table reports the mean of all variables by treatment and control, the differences of means and their standard errors adjusted for clustering using formulas in Liang and Zeger (1986). A * denotes significance level at 5% or 10%.

Table 11

The Treatment Effect on English and Math Bagrut Outcomes Estimated Using the Sharp Regression Discontinuity Sample

	Math			English		
	Attempted exams	Attempted credits	Awarded credits	Attempted exams	Attempted credits	Awarded credits
Control group mean	1.01	1.86	1.39	0.86	2.46	1.95
Control for correct matriculation rate	0.047 (0.030)	0.100 (0.056)	0.244 (0.078)	0.129 (0.031)	0.212 (0.081)	0.177 (0.104)
No control for correct matriculation rate	0.046 (0.032)	0.093 (0.058)	0.231 (0.079)	0.132 (0.030)	0.199 (0.079)	0.177 (0.108)

Note: The table reports treatment-control differences for the three outcomes for English and Math. Standard errors are adjusted for clustering at the school level using formulas in Liang and Zeger (1986) and presented in parenthesis. The sample include 4,087 students of year 2000 and 3,941 students of year 2001 (same samples of Table 10). Students level controls include the number of siblings, gender dummy, father's and mother's education, a dummy indicating an immigrant student, a set of dummy variables for ethnic background, a set of dummies for the number of credit units gained in the relevant subject before treatment, overall credit units gained before treatment and the average score in the relevant tests.

Table 12

Effects on English and Math Bagrut Outcomes by Quartiles of Previous Test Scores, the Sharp Regression Discontinuity Sample

	Estimates by quartile: Math				Estimates by quartile: English			
	1st quartile	2nd quartile	3rd quartile	4th quartile	1st quartile	2nd quartile	3rd quartile	4th quartile
Attempted exams								
Treatment effect	-0.199 (0.075)	0.192 (0.059)	0.144 (0.052)	0.095 (0.064)	0.156 (0.088)	0.166 (0.056)	0.060 (0.057)	0.132 (0.070)
Control group mean	0.552	0.967	1.142	1.322	0.670	1.060	0.981	0.708
Attempted credits								
Treatment effect	0.090 (0.124)	0.246 (0.114)	0.125 (0.087)	-0.068 (0.118)	0.313 (0.171)	0.414 (0.089)	-0.016 (0.130)	0.132 (0.179)
Control group mean	0.898	1.653	2.132	2.628	1.549	2.574	2.992	2.614
Awarded credits								
Treatment effect	0.165 (0.118)	0.361 (0.114)	0.391 (0.093)	0.060 (0.111)	0.375 (0.161)	0.435 (0.132)	-0.100 (0.137)	-0.030 (0.167)
Control group mean	0.406	1.019	1.633	2.372	0.886	1.877	2.553	2.347
Bagrut rate								
Treatment effect	0.044 (0.023)	0.089 (0.030)	0.036 (0.026)	-0.020 (0.034)	*	*	*	*
Control group mean	0.063	0.380	0.690	0.831	*	*	*	*

Note: The table reports treatment effects for Math and English outcomes. Treatment effects vary by quartile of summary Bagrut score through December 2000. Standard errors in parenthesis are adjusted for clustering at the school level using formulas in Liang and Zeger (1986). Sample identical to that of Table 11. All models control for the student's and school characteristics, lagged outcomes and also school fixed effects.

* The estimates for the Bagrut rate are the same for Math and English because it is the same outcome in an identical sample of students.

Table 13

Program Effect on Untreated Subjects Estimated with Various Methods

Estimation Method	History		Biology		All Untreated Subjects	
	Control Group Mean	Program Effect Estimate	Control Group Mean	Program Effect Estimate	Control Group Mean	Program Effect Estimate
OLS						
Attempted exams	0.447	0.161 (0.088)	0.722	0.161 (0.085)	3.736	0.205 (0.171)
Attempted credits	0.472	0.150 (0.087)	0.065	0.044 (0.032)	7.787	0.468 (0.242)
Awarded credits	0.303	0.027 (0.046)	0.029	0.016 (0.019)	4.230	0.054 (0.193)
Propensity Score Matching						
Attempted exams	0.345	0.158 (0.084)	0.573	0.156 (0.090)	3.643	0.165 (0.187)
Attempted credits	0.375	0.145 (0.084)	0.053	0.062 (0.034)	7.808	0.495 (0.252)
Awarded credits	0.229	0.020 (0.046)	0.029	0.024 (0.019)	3.947	0.013 (0.196)
RD - Natural Experiment						
Attempted exams	0.146	-0.084 (0.041)	0.629	0.111 (0.102)	2.786	0.258 (0.198)
Attempted credits	0.148	-0.066 (0.040)	0.011	0.251 (0.135)	6.216	0.063 (0.241)
Awarded credits	0.089	-0.077 (0.042)	0.008	0.136 (0.088)	2.688	0.282 (0.194)
Sharp RD						
Attempted exams	0.206	0.138 (0.084)	0.351	0.146 (0.089)	2.888	0.151 (0.205)
Attempted credits	0.206	0.160 (0.083)	0.018	-0.114 (0.146)	6.638	0.020 (0.184)
Awarded credits	0.144	0.064 (0.079)	0.003	-0.140 (0.075)	2.848	-0.059 (0.172)

Note: The table reports treatment-control differences for the three outcomes for History, Biology and all untreated subjects (including History and Biology). Standard errors are adjusted for clustering at the school level using formulas in Liang and Zeger (1986) and presented in parenthesis. The OLS sample is 72,378 (identical to the English sample in Table 2). The propensity score matching results are based on the PSM English sample (identical to the English sample in Table 3).

Table 14

The Effect of Pay For Performance on Teaching Methods and Teacher's Effort

	English teachers		Math teachers	
	Comparison group sample mean	Treatment-control difference	Comparison group sample mean	Treatment-control difference
Teaching methods:				
Teaching in small groups	0.596	0.122* (0.052)	0.665	0.007 (0.050)
Individualized instruction	0.559	0.148* (0.052)	0.645	-0.027 (0.046)
Tracking by ability	0.429	0.232* (0.053)	0.404	0.151* (0.048)
Adapting teaching methods to students ability	0.938	0.051 (0.020)	0.932	0.020 (0.023)
Teacher's effort:				
Added instruction time during the year	0.323	0.051 (0.052)	0.505	-0.038 (0.048)
Number of additional instructional hours [^]	4.882	0.229 (0.761)	4.984	1.666* (0.761)
Added instruction time in period before Bagrut exam	0.211	0.203* (0.049)	0.291	0.083* (0.045)
Teacher's additional effort targeted towards:				
All students	0.624	-0.193* (0.068)	0.708	-0.061 (0.050)
Weak students	0.306	0.081 (0.066)	0.245	-0.055 (0.044)
Average students	0.024	0.028 (0.027)	0.017	0.043* (0.020)
Strong students	0.000	0.029 (0.018)	0.000	0.023* (0.011)
Number of observations	335		427	

Note: Standard errors in parenthesis. Asterisks denote estimates which are significantly different from zero at 5% significance level. The English sample includes 141 of the 168 12th grade English teachers that participated in the program. The Math sample includes 169 of the 203 12th grade Math teachers that participated in the program.

[^]All the variables in the table are dummy indicators (that equal to 1 or 0) except the variables that are noted by a [^].

Table A1

Teacher's Education And Demographic Characteristics

	English teachers		Math teachers	
	Sample mean	Treatment-control difference	Sample mean	Treatment-control difference
Teacher demographics:				
Age	45.00	-0.697 (1.023)	44.20	0.301 (1.004)
Gender (Female=1)	0.81	-0.005 (0.044)	0.59	-0.024 (0.052)
Born abroad	0.62	-0.160* (0.053)	0.48	0.014 (0.053)
Teacher education:				
Teacher certificate	0.02	0.025 (0.016)	0.03	0.049 (0.019)
B.A in education	0.09	0.012 (0.031)	0.08	0.070 (0.028)
B.A	0.46	-0.066 (0.056)	0.41	0.024 (0.052)
M.A + Ph.d	0.43	0.034 (0.055)	0.47	-0.143 (0.052)
Teaching experience (years)	18.60	-1.470 (0.982)	19.01	-0.139 (1.009)
Education quality:				
Degree from top universities	0.18	0.089* (0.042)	0.20	0.040 (0.043)
Degree from other universities	0.33	0.004 (0.053)	0.33	0.048 (0.050)
Degree from teacher colleges	0.08	-0.020 (0.031)	0.10	-0.045 (0.032)
Degree from overseas universities	0.41	-0.067 (0.055)	0.36	-0.050 (0.051)
Number of observations	329		358	

Note: Standard errors in parenthesis. Asterisks denote estimates which are significantly different from zero at 5% significance level. The English sample includes 141 of the 168 12th grade English teachers that participated in the program. The Math sample includes 169 of the 203 12th grade Math teachers that participated in the program. Top universities: Hebrew University in Jerusalem, Tel-Aviv, Technion and Weizman Institute. Other Universities: Bar-Ilan, Ben Gurion and Haifa university.