# Paying for Performance: The Effect of Financial Incentives on Teachers' Effort and Students' Scholastic Outcomes[*]

Victor Lavy

The Hebrew University of Jerusalem and CEPR

May 2003

[*] Special thanks go to Alex Levkov for outstanding research assistance. I benefited from comments of seminar participants at EUI, Harvard, Hebrew University, MIT, NBER 2002 Summer Institute, Princeton, Tel-Aviv University, The Tinbergen Institute and the Public Economics Conference in Bristol. 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 Teachers' Financial Incentives on Students' Scholastic Outcomes

## Abstract

Performance-related pay for teachers is being introduced in many countries, but there is little evaluation to date on the effects of such programs. This paper evaluates a particular incentive experiment. The incentive program is a rank-order tournament among teachers of English, Hebrew, and mathematics. Teachers were rewarded with cash bonuses for improvements in their students' performance on high-school matriculation exams. Since the schools in the program were not selected at random, the evaluation is based on comparison groups. Three alternative identification strategies are used to estimate the causal effect of the program: a natural experiment stemming from measurement error in the assignment variable, a regression discontinuity method, and propensity score matching. The results of all three methods tell a consistent story: teachers' monetary performance incentives have a significant effect on students' achievements in English and math. No spillover effect on untreated subjects is evident and the general equilibrium impact of the program is positive as well. The program is also more cost-effective than alternative forms of intervention such as 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 teachers may be motivated by incentive pay. However, there is little evidence on the effect of teachers' incentives at schools. In this paper we present evidence from an experiment that offered bonus payments to teachers based on their class performance. 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' effort responds 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, designed to improve their students' achievements on matriculation exams in English, Hebrew, and Mathematics.

This paper evaluates an Israeli program where teachers were rewarded with cash bonuses for improvements in their students' performance on the high-school matriculation exams. The bonus program was structured as 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

[1] Examples of such programs in the USA are 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 to start 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; Conley and Odden, 1995; Elmore, Abelmann and Fuhrman, 1996; Kelley and Protsik, 1996 and Sadowski and Miller, 1996.

[2] The theory of Individual and group incentives in rank order tournaments are discussed by Lazear and Rosen (1981), Green and Stokey (1983) and Prendergast (1999).

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.

The main questions of interest in this experiment relate to the effect of the program on teachers effort and 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 the 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 is 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. Three alternative identification strategies were used to estimate the causal effect of the program. The first strategy is based on a measurement error in the assignment variable that was used to assign schools to the program. The assignment variable, the mean matriculation rate of the school in 1999, was compared with a given threshold (45 percent): any school with a rate equal or below the threshold was assigned to the program. The administrators of the program, unaware that the assignment variable used was measured erroneously, assigned some schools to the program mistakenly. Since the measurement error was random and unrelated to the potential outcome, as will be shown, assignment to the program was actually random in a sub-sample of the schools. This resulted in a natural experiment that could be used to identify the effect of the program in this sub-sample. This identification strategy is abetted by the use of panel data (before and after the program) that allow an estimation of differences in differences estimates in the "random" assignment sub-sample.

The second identification strategy is based on the assignment rule that determined program participation. This process was based on a threshold function of the school matriculation rate: schools with a rate equal to or lower than a critical value were included in the program; others were excluded.

Since the school matriculation rate varies from year to year, to some extent due to random effects, it is very likely that untreated schools that were just above the threshold resembled schools that were at or below the threshold. The narrower the threshold value of the band, the more likely such a similarity is. We exploited this “regression discontinuity“ feature of the assignment mechanism to estimate the effect of the program using the sample of schools within a relative narrow band around a sharply drawn threshold.

The third approach makes use of 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 method. Having various dimensions of lagged outcomes improve the likelihood that pupils are matched also on non observables attributes.

Section I of this paper provides background information about the Israeli school system, describes the teachers’ incentive program, and discusses the theoretical context for pay-for-performance programs. Section II discusses the evaluation strategy. Section III presents the three approaches used for causal identification of the program effect and the empirical results of the effect of teachers’ incentives on the mathematics and English performance of students exposed to treatment obtained with each approach. Section IV presents evidence on the effect of incentives on teachers’ behavior in the classroom during the program. Section V 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 and the classroom. The concluding section presents evidence on 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 presented in the paper suggest that student achievement improves significantly when teachers are offered financial incentives that reward this. These improvements correspond to changes in teachers’ behavior as a result of the program: greater teachers’ effort, changes in teaching methods and more teachers’ awareness and responsiveness to students’ needs. The results also suggest that it is difficult to predict who are the better teachers (those who eventually won financial awards) by conventional teacher characteristics such as age, gender, education, teaching certification, and years of teaching experience. However, some measures of teachers’ education quality, such as quality of college education, were positively correlated with teachers' quality.  Finally, the cost-benefit comparison of other relevant interventions suggests that individual teachers' financial incentives are more efficient than teachers' group incentives and equally efficient as paying students monetary bonuses to improve their performance. All three incentive programs were more efficient then 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 imposes additional risks on employees, which is costly to employers through higher wages (Prendergast, 1999). 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 an individual member; the group (of teachers) often has better information 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 exert greater effort 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.[3]

Tournaments as an incentive scheme has been 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 between individuals for those positions (Lazear and Rozen, 1982, Green and Stokey, 1983). All that matters for winning in such tournaments is not the absolute level of performance, but how well one does relative to others. Although promotion is not an important career feature among teachers, emphasize on relative rather then absolute performance measures is relevant for a teachers incentive scheme for two reasons. First, awards based on relative performance and a fixed set of prizes would stay within budget. Second, in a situation were there are no obvious standards that could be used, as a basis for absolute performance, relying on how well teachers do relative to others seems a preferred alternative. We therefore used the structure of a rank order tournament for the teachers incentive experiment described below.

### *2.2 The Israeli Secondary Schooling*

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 to all teachers in the winning schools. The purpose of the program was to improve students achievement on the bagrut (matriculation) examinations, a series 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: units) per subject.[4] Some subjects are mandatory and many must be taken at the level of three units at least. Tests that award more units are more difficult. A minimum of twenty units is required to qualify for a matriculation certificate. About

[3] 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.

[4] In Israel, the high school matriculation exam – known as the “Bagrut” – is a pre-requisite for admission to universities and is one of the most economically important education milestones. Similar high school matriculation exams are found in many countries and in some American states. Examples include the French Baccalaureate, the German Certificate of Maturity

52 percent of high-school seniors received matriculation certificates in 1999 and 2000, i.e., passed enough exams to be awarded twenty units by the time they graduated from high school or shortly thereafter (Israel Ministry of Education, 2001).

In early December 2000, the Ministry of Education announced 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 through twelve in advance of matriculation exams in these subjects in June 2001. In December 2000, the Ministry conducted an orientation with principals and administrators of the forty-nine schools. The program was announced as a voluntary three-year experiment.[5] All 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 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, schools must have a recent history of relatively poor performance in the Mathematics or English matriculation exams,[6] and the most recent school 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.[7]

The initial intention was to limit the program to math and English teachers. Under pressure from the teachers' union, teachers of Hebrew and Arabic were added. Schools were also allowed to replace the language (Hebrew and Arabic) teachers with teachers of other core matriculation subjects

(*Reifezeugnis*), the Italian Diploma di Maturità, the New York State Regents examinations, and the recently instituted Massachusetts Comprehensive Assessment System.

[5] 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 continued as planned for a second and third year.

[6] Performance was measured by the average passing rate in the math and English matriculation tests during the last four years (1996-99). Two occurrences or more of any of these rates being lower than 70 percent was considered a poor performance. English and math were chosen because they are the subjects with the higher failing rate among the matriculation subjects.

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

(Bible, literature, or civil studies). The schools and teachers were informed on December 20, 2000, about their participation in the program.

### *3.2 The Israeli Policy 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 stand for 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 or 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. Ranking was based on the difference between the actual outcome and a value predicted on the basis of a regression that controlled for the students' socioeconomic characteristics, the level of proficiency in each subject, and a school fixed 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 enrolment lists with itemizations 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 in 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 of 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.[8]

The program included 629 teachers, of whom 207 competed in English, 237 in mathematics, 148 in Hebrew or Arabic, and thirty-seven 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 teacher won two awards each; twelve math teachers won two awards each, and one Hebrew teacher won two first place awards totaling $15,000.

A follow-up survey of teachers in the program was conducted during the summer after the end of the school year. Seventy-four percent of teachers were interviewed and there were very few refusals. Most absences among potential interviewees were due to wrong phone numbers or teachers not being reached on the phone after several attempts were made.[9] 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 the school principal and the program coordinator—and 75 percent thought that the information was complete and satisfactory. Almost 70 percent of the teachers were familiar with the award criteria and about 60 percent of them thought they would be among the award winners. Only 30 percent did not believe they would win; the rest were certain about their chances. Two-thirds of the teachers thought that the incentive program would lead to an improvement in students' achievements.

## 3. Evaluation Strategy

The first evaluation issue to address is the non-random selection of schools and, therefore, of teachers for the program. Denoting by $Y_i^1$ the outcome of a pupil i who were exposed to teachers in the program and by $Y_i^0$ the outcome were a pupil was not exposed to teachers in the program. The impact of the intervention for the ith 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 that we want to estimate is the impact of

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

treatment on the treated, i.e. $E(Y_i^1 - Y_i^0 \mid T_i = 1)$, where $T$ is one for pupils in schools with treated teachers and zero in the schools not included in the program. What we do observe is $E(Y_i^1 \mid T_i = 1)$, which is the average outcome for pupils exposed to teachers who participated the bonus program. To construct the counterfactual $E(Y_i^0 \mid T_i = 1)$, we outline in the next section three strategies meant to surmount this difficulty and help to identify the causal effect of the program.

The evaluation may include English and math teachers only because school participation in Hebrew and Arabic was optional and all schools had the choice of replacing these subjects with other core matriculation subjects. Since some schools did exercise this option, the sample of schools that elected not to do so is endogenous.

A second issue of concern relates to the implications of the teachers' potential strategic behavior. In other words, the teachers' increased investment of time and effort, due to the incentives offered them, 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. By implication, we should estimate the effect of the program on the students' overall outcomes and not only on the treated subjects. For the sake of simplicity, let us separate students' educational outcomes into awarded subjects (*Y1* and other subjects (*Y2*). If students invest less time and effort in *Y1*, then the effect on *Y2* may be negative. However, the additional instruction time in the treated subjects may free some of the students' time for other subjects, so that the effect on *Y2* may be positive. Estimating the effect of treatment by examining the change in the rewarded subjects only may overstate or understate the treatment effect of the program, if there are indeed negative or positive spillover effects due to change in time and effort allocation.

We may address this spillover or substitution aspect of the evaluation by estimating the effect of the program on the outcomes of all other "untreated" subjects. However, the number of subjects that we may use as 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, as we will recall, schools were allowed to include in the program teachers of one other

[9] The survey was conducted during the summer break and therefore many teachers were away on vacation.

subject in lieu of Hebrew or Arabic. Some school exercised this option and included subjects such as Bible and social studies, which basically excluded these subjects from being considered untreated. Instead of estimating the effect of the program on each subject, we may generate a summary measurement that takes account of all the exams, e.g., the total number of credit units earned on all the exams. Another possibility is to confine the focus to untreated subjects that had the largest sample size. Using this criterion, two subjects, history and biology, stand out. A third alternative is to estimate the effect of the program on the students' matriculation status. This is an overall high school achievement measurement that encompasses the outcomes of all matriculation tests from tenth grade through the end of twelfth grade. Below we report results using all three alternatives: evidence of the effect of the program on the biology and history outcomes, evidence of the effect on total credit units accumulated during the program period in untreated subjects, and evidence of matriculation status.

A third issue to address is how to measure the outcome variables. In each of the treated subjects, the requirements often include several exams. For example, a student who takes mathematics at the proficiency level of three units has to take two exams, one for the first unit and the second for the other two units. In some subjects there are additional exams, such as a lab test in science subjects and an oral test in languages (English, Hebrew, or other). The final score is a weighted average of all components, very often, but not always, the weights reflecting the credit units of each component. Since the information about the exact weights was not available, I could not use the final score as an outcome. Instead I used in this study three different but related measures of outcomes for each subject: the number of tests taken by a student in the given subject, the total number of units in the tests attempted, and the total units earned (a measure that reflects the pass rate in each exam weighted by its number of credits). The second and third measurements reflect the proficiency level of the curriculum. Below we estimate the effect of the program on these three outcomes in every treated subject and in biology and history. We will also use as an outcome the total number of units accumulated by each student in all untreated subjects.

## 4. Identification Strategies, Estimation and Results

### *4.1 Natural Experiment: 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 they do every year.[10] As a result, the matriculation rates used for assignment to the program were inaccurate in a majority of schools. The measurement error is useful for identification of the program effect. In particular, conditional on the true matriculation rate, program status is 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.[11] Most (80 percent) measurement errors were negative, 17 percent were positive and the rest had no 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 suggests that the variance of the measurement error is lower at low matriculation rates, this is most likely due to the floor effect that bounds the size of the negative errors: the lower the matriculation rate, the lower the absolute maximum size of the negative errors. Similar evidence arises when the sample is limited to the ninety-seven schools that were eligible for treatment, those from which forty-nine schools were assigned for treatment (Figures 3 and

[10] Matriculation status depends on the fulfillment of many requirements (e.g., a minimum number of credit units and the coverage of compulsory subjects such as math, foreign language, and Bible) that tend to vary by school type (technical, agricultural or regular, Jewish or Arab, religious or nonreligious) 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 student file that includes the results of the matriculation exams becomes available in October of every year (for the cohort that graduated in June of that year). However, it is updated continuously and the final version is not completed until late December of the same year.

[11] The sample was limited to schools with positive (> 5%) matriculation rates.

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 1 presents the estimated coefficients from regression of the measurement error on student characteristics (mother and father year of schooling, number of siblings, gender and immigrant status), lagged students' outcomes (in treated and untreated subjects and also on total lagged credits earned and the respective average score), and school characteristics (indicators of whether it is a religious or secular school and whether it is an Arab or Jewish school). These regressions were run with school level means of all variables, separately for the whole sample (507 high schools) and only the eligible sample (97 schools). The whole set of regressions were estimated twice, once with the 2000 data and once with the 2001 data.

The first panel of Table 1 presents 20 estimated coefficients from regressions of the 1999 measurement error on student's characteristics, only one of which 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 24 estimated coefficients from regressions of the 1999 measurement error on student's pre-program outcomes, only four of which are marginally significantly different from zero (English and history lagged credits in the eligible school sample in year 2000 sample, the average score in the all school sample of year 2001 and the English lagged credits in the eligible school year 2001 sample). Based on these results we can conclude that the 1999 measurement error is uncorrelated with observable characteristics that are likely be correlated with potential outcomes.[12]

The identification strategy 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 matriculation rate for 1999 and ε the measurement error. Let $T$ denote 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 sub-sample 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 amount to matching schools on the basis of the value of $S^*$. Figure 6 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 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. We should bear this in mind when interpreting the results, especially in the case of treatment heterogeneity. However, it is important to note that the range of the matriculation rate in this sample year 2000 is much wider, both samples of participating and non-participating schools, from 32 to 79 percent. This wider range of the distribution of the school mean matriculation rate mitigates somewhat the limitation in terms of external validity of the findings.

Table 2 presents the pre-program (2000) and post-program (2001) means of students (those graduating twelfth grade) and school characteristics for the seventeen treated schools and the twelve control schools. The student background characteristics include father's and mother's schooling,

[12] Another possible way to test for a random nature of the measurement error was to test if it is serially uncorrelated. However, we do not have more than one year of data on the initial matriculation rate and its revised value, and so cannot

number of siblings, and dummies for gender and for immigration status. 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 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 2 presents students' outcomes in the form of units earned before twelfth grade in treated subjects (English, math) and untreated subjects (biology, history). For twelfth graders in the treatment year (2001), this measure reflects pre-program (or lagged) 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 units. 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 2 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 1 suggests that, generally speaking, the treatment and control schools are well balanced in most pupils and school characteristics, reflecting the basic similarity of the two sets of schools. These findings support the notion of a natural experiment as a strategy to identify the effect of the teacher bonuses program in this sample of treated schools. Nevertheless, it is still necessary to control for all these variables in the estimation to net out the effect of any remaining differences. In particular, we should control for the true 1999 matriculation rate. The evidence about the effect of the program will be based on regressions for a sample that includes twelfth graders in 2000 and 2001 (stacked panel data). The explanatory variables will include pre-program outcomes, pupil and school covariates, and constant school effects. The estimated treatment effect will be equivalent to differences in differences estimate embedded in a natural experiment setting. The

compute the measurement error for any previous years.

constant school effects will absorb any remaining permanent differences, observed and unobserved, between the treated and the control schools. The estimation framework is discussed below at greater length.

*Estimation*

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

(1) $$Y_{ijt} = \alpha_j + X_{ijt}{}' \beta + Z_{jt}{}' \gamma + \delta\, T_{ijt} + \varepsilon_{ijt}$$

where *i* indexes pupils; *j* indexes schools; *t* indexes years 2000 and 2001, and *T* is the assigned treatment status. The model includes a vector of student-level covariates *X*, and a vector of school-level covariates *Z*. We estimate the regressions on the basis of pooled data from both years (the two adjacent cohorts), stacked as school panel data with school fixed effects ($\Phi_j$) included in the regression:

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

where j indexes schools. This model also includes a constant effect for each year ($D_t$) with a factor loading $\eta$. 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 (2) is 1 for treated schools in year 2001 and 0 otherwise). The estimated treatment effect in this model is a difference in differences estimate. Its advantage is that it nets out any correlation between the outcome variable and any school characteristic that did not change between 2000 and 2001. In the next section we present the results obtained from the estimation of this model with the measurement-error sample.

*Results*

The first panel rows in Table 3 present the evidence from regressions using the stack panel data for a sample of twelfth-grade students in 2000 and 2001. The treatment indicator equals 1 for year-2001 students in the seventeen treated schools. The regressions include, as controls, individual covariates (gender, father’s and mother’s education in terms of years of schooling, ethnicity, and subject-specific and total matriculation units earned before treatment), school covariates (a dummy for religious schools and a dummy for Arab schools and school fixed effects), and a dummy variable for year 2001.

The treatment effect is estimated with two alternative specifications, with and without the correct 1999 matriculation rate included as a control. Treatment effect estimates are presented for three outcomes: number of exams attempted, number of units attempted and number of units earned, all three outcomes in English and math separately. The standard errors reported in the table are adjusted for clustering, using formulas set forth in Liang and Zeger (1986).[13]

The treatment effect in English and math, for all outcomes, are positive but vary 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. There is more room for improvement in units earned than in the first two outcomes because units earned can be increased by raising the score from fail to pass in any given test, while the other outcomes may be improved by taking more exams or changing the curriculum which is difficult to do in the middle of the year. Indeed, the estimated relative effects on units earned are greater than on the other two outcomes. The effect of treatment on units earned in math is 0.256, a 19 percent improvement relative to the mean of the control schools (1.35). The effect of treatment on awarded credit units in English is 0.361, a 21 percent improvement relative to the mean of the control schools (1.7). The relative improvement in units attempted is much lower—6.7 percent in math and 9.7 percent in English. These results imply that the effect of teachers' incentives works through two channels. The first channel is the increase among treated students of the attempt rate of exams and units. The second channel is the increase in the probability of passing each exam successfully. However, the evidence in this sample suggest that the second channel is much more important in the overall effect of the program.

The interpretation of the above results as causal is based on the program status being randomly assigned by the measurement error, conditional on the true 1999-matriculation rate. Indeed

[13] Liang and Zeger (1986) developed a Generalized Estimating Equation (GEE) framework, which allows for an unrestricted correlation error structure and can be used for binary outcomes as well. The disadvantage is that the validity of GEE inference turns on an asymptotic argument based on the number of clusters. Our sample of 30 schools may be considered too small for asymptotic formulas to provide accurate approximation to the finite-sample sampling distribution (Thornquist and

the treatment effect estimates are sensitive to the exclusion of the correct 1999 matriculation rate as a control. For example, without this control the treatment estimated effect on math awarded credits is much lower, 0.163 versus 0.256, and it is also less precisely estimated. The English treatment estimate is also lower but only marginally.

Table 3, lower panel, presents also the treatment effect estimates for untreated subjects, biology and history. None of the estimates is significantly different from zero at the 5% level of significance except the effect of attempted exams in biology. In fact, the estimates in all three outcomes are negative in history and positive in biology, but practically speaking, given the large estimated standard errors, these point estimates reflect most likely chance deviations from zero. We also estimated the treatment effect on all units attempted and earned other than those in the treated subjects. The effect on total units attempted was 0.063 (S.E.= 0.318); the effect on total units earned was 0.282 (S.E.=0.257). Both estimates are not significantly different from zero. We think of these results as evidence that there were no spillover effects of the math, English, and languish teachers' incentives on other outcomes in other matriculation exams.

To assess the overall effect of the program, taking into account its effect on all treated subjects and untreated subjects, we also estimated the treatment effect on the matriculation rate. This estimate is 0.019 (S.E.=0.011). This effect of a 1.9 percent increase in the matriculation rate is marginally significant and it represents a 5 percent improvement relative to the mean of the control schools in this sample.

### *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 3, I estimated models allowing treatment effects to vary with lagged outcomes. In particular, I allowed for an interaction of treatment effect with the mean credit-unit-weighted average score on all previous matriculation tests (coding zeros for those with no tests). Using this average score, which is a powerful predictor of students' success in the math and English tests, I coded dummies for each

Anderson; 1992). However, since we are using 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 60.

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:

$$(3) \qquad Y_{ijt} = \alpha + X_{ijt}'\beta + Z_{jt}'\gamma + \Sigma_q d_{qi}\mu_q + \Sigma_q \delta_q T_{jt} + \Phi_j + \eta D_t + \varepsilon_{ijt},$$

where $\delta_q$ is a quartile-specific treatment effect and $\mu_q$ 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, a claim supported by the fact that the mean ‘Bagrut’ rate in this quartile in year 2000 was 90 percent. We should therefore expect to find no effect of the teachers’ incentive program on students in this quartile. On the other hand, those students with scores around or below the mean of the score distribution are in a range that extra effort of their teachers and of themselves may have made a difference. We therefore look for significant estimates for students mainly in the first to the third quartile.

Table 3a reports results of the estimation of equation (3), using the ‘measurement error’ sample’ for the three math and English outcomes and also for the matriculation rate. Significant positive effects on number of exams and credits attempted are estimated only for students in the first and second quartile. The zero effect on these outcomes in the third and fourth quartile are not surprising since all students in these quartiles are expected to take all exams as scheduled. This pattern is equally evident for math and English. The quartile pattern of the effect on passing rate in the exams (awarded units) reveals a significant positive effect also in the third quartile in math but not in English. The largest absolute effect on awarded credit units is in the second quartile. The effect on math is an increase of a half a unit against a mean of one unit in the control group, an impressive 50 percent increase. In English the effect in the second quartile is an increase of 0.58 units against a mean of 2 units in the control group, implying an increase of 30 percent due to the program. However, the most dramatic effects on awarded credits is the first quartile: in math 74% (a 0.258 increase against the control group mean of 0.347) and in English 78% (a 0.707 increase against the control group mean of 0.911).

The last panel in Table 3a 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% counterfactual, the mean of the second quartile of the control group. We find no effect on the matriculation rate in the first quartile, an expected result because most of these students lack many other requirements needed to qualify for matriculation. Only about 5 percent of the first quartile students gain matriculation and the program did not change this rate.

### *4.2 Identification based on Discontinuity in the Assignment Variable*

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 discontinuity in the treatment assignment mechanism may be exploited as a second source of exogenous identification information for evaluation of the effects of the teachers' bonuses program.[14] 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 non-participants (in this context the term marginal refers to those schools not too far from the threshold for selection). 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 can be presented as follows. Let r be the threshold for participation, so that (r=45 or r=43) I=1(S $\leq$ r). The participation status for schools in a neighbourhood of r changes for non-behavioural reasons. Marginally participant ($r^-$) and marginally non-participant ($r^+$) schools define "quasi-experimental" groups. The main drawback of this approach is that it allows estimating the effect only for marginally exposed schools. In the presence of heterogeneous impacts it only permits to identify the mean impact of the intervention at the threshold for selection, which might be different from he effect for schools away from the threshold for selection.

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. The 0.40–0.52 range may be too large, but we can control for the value of the assignment variable (the mean matriculation rate) in the analysis. We should also note that there is some overlap between this sample and the measurement-error sample: nine of the fourteen treated schools and five of the twelve control schools belong to the group of treatment and control schools, respectively, of the measurement-error sample. However, we note that 12 of the 26 schools (almost 50%) included in the discontinuity sample were not among 29 schools that make the measurement error sample. This suggest that the overlap between the two sample still leaves enough "informational value added" in each of the samples.

Table 4 presents descriptive statistics for the treated and control groups in the discontinuity sample for the cohort that graduated in the year before treatment (2000 seniors) and for the treatment cohort (2001 seniors). The treatment-control differences and standard errors in the background variables (Columns 3 and 6) reveal that the two groups are very similar in both years in all background 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 year 2001.

The second panel in Table 4 shows that before the program began students in treated schools earned fewer math units than students in control schools. This gap is evident in both years: the treatment-control difference is –0.339 in 2000 and –0.312 in 2001. The difference in 2000 is significantly different from zero; the difference in 2001 is marginally so. A discrepancy in the same direction, although not significant in either year, is observed in total lagged units. In English, however, the opposite is observed: a positive and significant difference in English units in favor of the

[14] Regression discontinuity designs are 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).

treated schools in both years, although the difference is significant only in 2000. No significant treatment-control differences are observed in biology or history.

The third panel in Table 4 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, we should expect the treatment-control difference in the correct matriculation rate to be negative as well. The two differences are of similar magnitudes—0.061 and –0.054—and both had low standard errors.

The evidence presented in Table 4 suggests that the treatment and control schools are balanced in some individual and school characteristics but differ in pre-program matriculation rates and in the achievement outcomes in the treated subjects. Although the differences are small, they should be controlled for directly in the empirical analysis. Since the measured differences may reflect other unmeasured differences, identification based on the discontinuity approach depends more than in the case of the natural experiment (measurement error) on the school constant effects model (equation 2) for estimating the treatment effect, since this model accounts for all unobserved but fixed correlates of potential outcomes. Controlling for students lagged outcomes, as we do in each model estimated, also increase the likelihood that all confounding factors are netted out.

### *Results*

The discontinuity sample was used to estimate models identical to those estimated with the measurement-error sample. In principle, the identification based on the regression discontinuity is conditioned on controlling for the erroneously measured matriculation rate that was actually used to assign schools to the program. However, the school fixed effects, which are included in each regression, controls for the 1999 measured with error rate and therefore the later should not be included as a control.

Contrary to the estimates based on the measurement error sample, there is no reason to expect the results based on the regression discontinuity sample to be sensitive to the control for the lagged true matriculation rate. However, for purpose of comparison I again  estimated two specifications, with and without controlling for the correct lagged (1999 and 2000) matriculation rates even though identification is not conditioned on this variable.

The results are presented in Table 5. The treatment effect estimates are very similar, qualitatively and quantitatively, to the results obtained using the measurement error/natural experiment approach and sample. The treatment effect estimates are positive and significantly different from zero for all English and math outcomes except for the math attempted exams outcome, which is only marginally significant. The estimated effect on earned units in math is 0.244 (S.E=0.078), just slightly lower in size and precision to the estimate obtained with the measurement-error sample. The estimated effect on English units earned is 0.177 (S.E. = 0.104), about half the estimate derived from the measurement-error sample and less precisely estimated. The estimates based on the regression discontinuity sample are expected to be downward biased because the control group schools have on average higher pre-program outcomes. This might explain why the discontinuity results are somewhat lower in comparison to the estimates derived from the measurement error sample.

We should note that the treatment estimates in Table 5 are not sensitive at all to the exclusion of the true lagged (1999 and 2000) matriculation rate as a control variable; the coefficients in the second row of Table 5 are practically identical to those presented in the first row. This result suggests that the sensitivity of the results in the measurement error sample (Table 3) to the control of the lagged correct matriculation rate was unique to that sample, which strengthens the credibility of the causal interpretation of the results.

The evidence in the second panel of Table 5 indicates that treatment has no effect on history and biology outcomes in the discontinuity sample as well except for the effect on Biology awarded credits, which are marginally significant.  The coefficients are all positive in history and all negative in biology, an opposite pattern to the one revealed in this respect in Table 3. We also estimated the effect of treatment on all units attempted and earned other than those in the treated subjects. The effect on total units attempted was 0.286 (S.E.= 0.334); the effect on total units earned was 0.166 (S.E.=.283). Neither estimate is significantly different from zero. The estimated treatment effect on the matriculation rate in this sample is 0.023 (S.E.= 0.012). These results may be viewed as supporting evidence for the causal interpretation of the estimated effects on math and English outcomes, no spillover effect on untreated subjects and overall positive effect on matriculation achievements of students.

Table 5a reports results of the estimation of equation (3), 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 3a: 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 Table3a and in the results about the effect on the English outcomes are somewhat lower. The effect on the matriculation rate of students is significant only in the second quartile, an increase of 8.9% which, is not much different from the 7.6% effect shown in Table 3a.

*4.3 Matching based on Observable Characteristics and Lagged Outcomes*

The third method we use for identification is matching. It is based on the assumption that one may account for all differences between treated and untreated subjects by controlling for observable characteristics. Matching may be implemented non-parametrically by defining cells using discrete characteristics.[15] The more characteristics, 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 the “propensity score”. To construct the counterfactual $E(Y_i^0 \mid P_i = 1)$ we assume that:

$$E(Y_i^0 \mid T_i = 1, X_i, Z_j) = E(Y_i^0 \mid T_i = 0, X_i, Z_j)$$

which means that given the observable characteristics of students ($X_i$) and schools ($Z_j$), the allocation to treatment and control is random. Under this assumption it is now well known (see Rosenbaum and Rubin, 1983) that:

[15] For an application of this method, see Angrist (1998).

$$E(Y_i^0 \mid T_i = 1, \Pr(T_i = 1, \mid X_i, Z_j)) = E(Y_i^0 \mid T_i = 0, \Pr(P_i = 1, \mid X_i, Z_j))$$

where $\Pr(T_i = 1, \mid X_i, Z_j)$ is the propensity score and is simply the probability of being assigned to treatment given observed characteristics. It follows that we can estimate the counterfactual by the sample analog of

$$E(Y_i^0 \mid T_i = 1) = E_{F^1}[E(Y_i^0 \mid T_i = 0, \Pr(T_i = 1, \mid X_i, Z_j))],$$

where $E_{F^1}$ denotes an expectation with respect to the distribution of the propensity score in the treatment sample.

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. We restricted the control sample to include only those observations whose value of the propensity score is within the range of the propensity score in the treatment sample. Imposing this common support condition in the estimation of the propensity score improve the quality of he matches and avoid a major source of bias (see Heckman, Ichimura and Todd, 1997). The next step is to split the sample in k equally spaced intervals of the propensity score. Within each interval, the average propensity score of treated and control pupils should not differ. In the current application this requirements was met by defining 100 intervals for the propensity score distribution. Students are then matched according to their propensity score by the *Nearest Neighbor Matching* method within the 100 intervals. Students for whom no suitable match could be found were dropped to ensure that the comparisons between treated and control students take a place over a range of characteristics in which suitable comparison do exist Within each of these cells the means of each characteristics did not differ between treated and control pupils.

Matching by propensity score allows us to trade off different observable characteristics against each another according to their importance, in order to find the best match for matched students among the untreated students. A good match on observable characteristics does not necessarily ensure a good match on important unobservables such as ability and motivation. Our unusually rich data, however, include many lagged achievement outcomes that, if included in the propensity equation, may improve the match on important unobserved individual attributes because they are likely to correlate. Therefore, our propensity equation uses individual and school characteristics and all lagged matriculation outcomes. In estimating the treatment effect by the matching method, we derive corrected standard errors by using numerical bootstrapping methods.

The advantage of the matching method over the first two methods described above is that it uses all forty-nine treated schools and searches for matches in all other 520 high schools countrywide. Before discussing the results of the match, however, we should address an important point. About 82 percent of twelfth-grade students in treated schools were enrolled in math classes during the treatment year and about 88 percent were enrolled in English classes. Not being enrolled in either subject means one of two alternatives: either the student completed his or her matriculation studies and exams in English or math by mid-semester of the senior year (by December 2000, before the program began), or the student dropped out of English or math class before December 2000.[16] The data do not allow us to distinguish reliably the two alternatives. However, each alternative has different implications for the measurement of lagged outcomes. To avoid these complication and to focus on treated students, we can include as treated students only those who were enrolled in English or math classes, respectively, and search for their matches in all other schools countrywide. Alternatively, we can search for matches for all students irrespective of their enrollment in English or math. Therefore, in comparing the results of the matching method with those of the other two methods, we should bear in mind that the estimation samples we used in the two last-mentioned approaches included all twelfth graders irrespective of their enrollment in English or math.

Table 5 presents descriptive statistics for the treated and control groups in the propensity score sample that is based on students enrolled in English and math classes separately. The English and math samples are different, reflecting the fact that the number of students enrolled in the English and math classes during the experiment were different. All variables that appear in the table were used in the matching equation. Matches were found for almost all treated students (95 percent) for both the

[16] A third possibility is that the student was erroneously deleted from the English and math class rosters that the schools submitted before the program began. This possibility can be neither verified nor excluded.

math and English samples, from 330 schools in the English sample and 350 schools in the math sample.

The first panel in Table 5 shows that the matching leads to a perfect match in all demographic variables in the English and math samples except for immigrant status. None of treatment-control differences in the background variables is statistically different from zero except for immigrant status, as noted.

The second panel in Table 5 shows that the treated students and their matched samples in both subjects are perfectly balanced in all six measures of lagged outcomes. This is evident not only in each of the specific treated subjects, English and math, but also in the lagged outcomes of the untreated subjects, history and biology, and the lagged total units and mean score in all untreated subjects. These results concerning lagged outcomes reinforce our confidence that the two samples are well balanced in terms of unobserved student covariates.[17] The third panel in Table 5 reveals no statistically significant treatment-control differences in the school covariates that were used in the propensity score equation.

Overall, the evidence presented in Table 5 suggests that the treatment populations and their matched samples are well balanced in all dimensions of comparison. The only covariate that was not perfectly balanced was immigrant status. In estimating the treatment effect, we will control for this variable as well as for all other variables reported in Table 5.

Table 5A presents the comparison between the treatment and the control group that is based on matching for all students irrespective of enrollment in English or math classes. The results are very similar to those presented in table 5: the two samples are well balanced in all variables except the immigration status.

The next section present results based on the estimation of equation (1) for year 2001, using the matched samples described above. The model with school fixed effects (equation 2) cannot be

[17] As another check of the quality of matching, I re-estimated the propensity score model leaving out from the equation all lagged outcome but the math and English lagged credits. I then checked how well balanced in terms of the left out lagged outcomes variables are the treated sample and its comparison counterpart. None of the mean differences of these outcomes (history and biology credits, total credits, average score) were significantly different from zero.,

estimated because the matching method is based on data of a single cohort, that of the treatment year only.

*Results*

Table 7 presents the results of estimating the treatment effect with the propensity score matched sample of the students enrolled in English and math classes.[18] The treatment effect estimates in English and math are positive; all are significantly different from zero and resemble in magnitude the estimated treatment effects obtained by use of the other two methods, as reported in Tables 4 and 5. Focusing for comparison on the effect of treatment on credits earned, the estimate for math in the match sample is 0.250, almost identical to the estimated effect in the measurement-error sample (0.256) and the discontinuity sample (0.244). The effect on English credits earned is 0.193 in the matched sample, equal to the estimated effect in the discontinuity sample (0.177) but smaller than the estimate in the measurement-error sample (0.361). However, the estimated effect on attempted units in English is identical in all three methods, at about 0.22. The relative size of the effects on attempt and earned credits reconfirm our earlier conclusion that the incentive programs affected both the rate of students who attempted to pass the math and English exams, but also, and even with a larger effect, it led to an increase in the passing rate of these exams.

The evidence in the second panel of Table 7 indicates that in the matched sample of treated students, treatment has no significant effect on any of the history or biology outcomes. For example, the estimated treatment effect is 0.028 (S.E. = .053) on history units earned and 0.018 (S.E. = .020) on biology units. The effects on attempted exams and units in history and biology are estimated more precisely but are still not significantly different from zero at a conventional significance level. The effect on total untreated units attempted or earned, although not shown in Table 7, is also not significantly different from zero. The treatment effect on the matriculation rate in this sample is 0.032 (S.E.=.013).

[18] The standard errors were estimated using boots strapping techniques. To account for clustering in the error term, we used a procedure that included in each round of estimation a random draw of samples of students (treatment and control separately) as well as a random draw of schools (treatment and control separately). In terms of the asymptotic bias in the estimates of the

The lower rows in the first panel in Table 7 presents the evidence from the matched sample that includes all students irrespective of their enrollment in English and math classes. The results are qualitatively similar to those obtained from the sample of students enrolled in English and math classes though the treatment effect estimates are somewhat lower. For example, for math earned credits the estimate in the all students sample is 0.162 and in the treated sample it is 0.250. However, when the point estimates are weighted by the proportion of students who were treated (0.88 in English and 0.8 in math), the differences between the two sets of estimates obtained from the two samples are almost eliminated. Further, when converted into relative effect size effects, the two set of estimates reflect similar size effects because the mean of awarded math and English credits is lower in the all students sample in comparison to the treated students sample (for math they are 1.60 versus 1.82 in the two samples, respectively).

Table 7a reports results of the estimation of equation (3), which allows treatment to vary by quartile of the distribution of the average score in pre-program matriculation exams, using the 'matched' sample of all students (the same sample referred in Table 7 as 'all students'). The pattern in the table is qualitatively very similar to that of Table 3a and Table 5a: almost no significant effects are estimated for the above average students (quartiles 3 and 4). However, one difference from the respective results reported in Table3a and Table5a is that the estimated effect on attempted exams and attempted credits are largest for the first quartile and the effect on awarded credits is equal for the first two quartiles. We should also note that the quartile pattern of the program effect on the matriculation rate is identical to that obtained from the other two methods of identification: a 7.5 % increase for students in the second quartile and no significant effect in the other quartiles.

## 6. Financial Incentives and Teachers' Effort

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 changes in teachers' behavior as a result of the program? Do they reflect

standard errors, the matching on the propensity score has an advantage over the previous two identification methods we used since the number of clusters is much larger.

greater effort on teachers' part or changes in teaching methods due to the program? To address these questions, a telephone survey was conducted among the English and math teachers who participated in the program. For comparison purposes, a similar survey was conducted with a similar number of nonparticipating English and math teachers. The comparison group was chosen for practical and logistical reasons and not necessarily because it was an appropriate comparison group vis-a-vis the treated schools. However, as Table A1 in the Appendix shows, the characteristics of the teachers in the two groups are very similar. For example, the treated and control teachers of English are identical in age (forty-five), gender (81 percent female), and schooling. Similar results are observed for math teachers.

Table 8 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 or strong students. Before turning to the evidence, we should admit the possibility that they being aware that they were part of an experiment may have affected teachers' responses to these questions. To minimize such a "Hawthorne" type bias, the survey was presented to interviewees (from treatment and control groups) 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 (survey).

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, expressed in the form of time devoted to student instruction beyond regular classroom time, especially in the weeks before the matriculation exams.

The mean of English teachers who taught students in small groups is 63 percent in the sample and 8.5 percentage points higher among teachers who participated in the program. Fifty-eight percent of teachers in the sample used individualized instruction, as against 69 percent of teachers in the program. Program teachers used tracking by ability much more than the comparison group of teachers, 42 percent versus 62 percent. Ninety-three percent of the control teachers reported having adapted their teaching methods to their students' ability; 100 percent of the treated teachers so reported.

Among math teachers, the behavioral change focuses on additional instruction time as opposed to teaching methods. The only difference in math teaching methods is in the prevalence of tracking by ability: 53 percent of program teachers as against 40 percent of other teachers. Most math teachers apparently invest additional time in teaching, beyond regular scheduled classroom hours, throughout the year. However, treated teachers added much more—4.8 hours per week as against 2.7 hours by other teachers, almost double the effort. The induced and intensified effort of treated math teachers is even greater as the matriculation exam period approaches: treated teachers begin special preparations with their students seven weeks before the exam date as against 4.5 weeks in the case of other teachers. English teachers who participated in the program also made an additional effort, but in only one dimension: preparing students for matriculation exams. Fifty percent of them report making such an effort, as against 37 percent of teachers in the comparison group.

Most teachers target the additional effort to all their students and to their weakest students. However, targeting of effort to weakest students is more prevalent among English teachers: 33 percent of those in the program versus 20 percent of other teachers. Among program math teachers, we see more effort targeted toward average students.

The evidence that the program led to significant increase in teachers' effort in the form of additional time of instruction, and the pattern indicating that it is directed mainly to weak and average students, coupled with the finding that teachers changed, sometimes even dramatically, their pedagogy and teaching methods, reduce very much the likelihood that the improvement in math and English matriculation outcomes are due to 'teaching to the test' phenomena. The observed real change in teaching technology and teachers' effort coupled with the increase in the matriculation rate, signal that these gains reflect real human capital accumulation.

## 7. Does Tournament Ranking Correlate with Teachers' Characteristics?

The results presented in this paper about the effect of teachers' incentives on students' achievements prompt us to conclude that individual teachers matter in improving schooling quality. Can we predict who the better teachers in our sample would be by some conventional measure of teacher quality? The ranking of teachers in the tournament was based on their students' average residuals, as determined from a regression of students and class characteristics on test scores. Teachers with positive residuals in both mean score and pass/fail regressions won awards. We can try to characterize the good teachers by seeking correlations between conventional teacher characteristics (shown in Table A1) and the teachers' residuals. We examined these correlations within a regression framework for English and math teachers separately.

The results support the view that we do not know how to measure teaching quality on the basis of conventional teacher characteristics such as age, gender, education, teaching certification, and years of teaching experience.[19] None of these variables was highly significant in the achievement residual regressions. Other variables, however, evinced significant correlations in the regressions. Being born and educated outside of Israel has 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 earned a higher academic degree were much more effective than other teachers. No similar effect was found for father's education.

We also correlated these teachers' attributes with the three measures of teachers' effort discussed in the previous section: whether the teacher added instruction time beyond regular classes during the program, how many weeks before matriculation exams the teacher exerted special effort, and how many hours of instruction per week he or she added during that time. No significant correlations were found between the personal attributes of participating teachers, in either subject, or these measures of effort.

## 8. A Cost Benefit Comparison to Alternative Interventions

How effective is incentive intervention relative to other forms of intervention that are also meant to improve matriculation results? We may compare three interventions in terms of their effects on the matriculation rate and their per-student cost. This comparison should be treated as an initial approximation and not as a set of exact figures. We should also bear in mind that the student populations treated in these programs were different, a fact that diminishes the validity of the comparison. The teacher incentive program cost $170 per student and led to a 3.1 percentage point increase in the matriculation rate, from 42 percent to 45.1 percent. The student bonus program evaluated by Angrist and Lavy (2002) cost $300 per student and elevated the matriculation rate by 6–8 percentage points, from 19 percent to about 26 percent. The group incentive program analyzed by Lavy (2002) cost $270 per student and boosted the matriculation rate by 1-2 percentage points, from about 45 percent to 47 percent. Another intervention relevant for comparison is the "Bagrut 2001 program" that the Ministry of Education initiated since year 2000. This intervention targets additional instruction time in small groups (2-6 students) in several matriculation subjects to weak students. The ministry evaluation results show that the program led to a 11-percentage point increase in the matriculation rate of the treated students at an average cost of $1,100 per student (Ministry of Education, Evaluation Division, May 2002).

Among the three incentive programs, the student bonus program was the most expensive in per-student terms but it is marginally more effective in cost-equivalence terms then 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. The added and targeted instruction time program led to a sharp increase in the matriculation rate but at a high cost, almost half the annual expenditure per student.[20] In terms of cost-equivalence it was the least effective among the four programs compared, similar to the teachers group incentive program.

Another way to benchmark costs and benefits is by comparison of the program cost per student to the likely economic benefits of the improved outcomes, for example of achieving a matriculation certificate. Against the cost of about $170 per treated student that led to a 3.1 percentage point increase in the matriculation rate we can compare the economic benefit of having a matriculation certificate. Angrist and Lavy (2002) estimate the economic benefit to an individual with 12 years of schooling of having a matriculation certificate at $4,025 per year. Given that the teachers

[19] See Heckman (2002) and Hanushek (2002) for discussion of this point.
[20] The 2001 average expenditure per student in regular high schools in Israel was about $2,200.

bonus experiment raised the mean probability of matriculation among treated students by 3.3 percent, it should increase annual earnings of treated pupils by 4,025 × .031 = $125 per person per year, allowing for the program cost to be recovered quickly, just after a year and a half.

Finally, we should note that teachers' incentives, beyond affecting motivation, might also have an impact, in the long run, through sorting and selection of teachers (Lazear 200 and 2001). Pay for performance will result in higher pay for the better teachers, which might 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 run effect is not feasible in this study.

## 9. Conclusions

The evidence presented in this paper indicates clearly that pay-for-performance incentives "work" among school teachers as well as in 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 not the outcome of the inputs of a single teacher but the product of the joint contributions of many teachers. The magnitude of the estimated effects and the evidence concerning teachers' differential efforts under an incentive regime suggest that teachers' incentives are a very promising path toward the improvement of school quality. The evidence culled from this new experiment adds important evidence to the results concerning group school incentives, presented in Lavy (2002).

## 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). “Achievements Awards for High school Matriculation: Research Methods and Findings.” Draft, June.

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.

Conley, Sharon and Odden, Allen. (1995). "Linking Teacher Compensation to Teacher Career Development*." Education Evaluation and Policy Analysis*, Summer 1995. 219-237.

Dehejia, Rajeev H. and S. Wahba. (1998). “Causal Effects in Non-Experimental Studies: Re-evaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association*, 2000 .

Elmore R. F, C. H. Abelmann 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.

Green, Jerry and Nancy L. Stokey (1983). “ A Comparison of Tournaments and Contracts.” *Journal of Political Economy 9*1: 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). “Publically 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-yee, and Scott L. Zeger, “Longitudinal Data Analysis Using Gerealized Linear Models,” B*iometrika* 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”, *Biometrica*, 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.

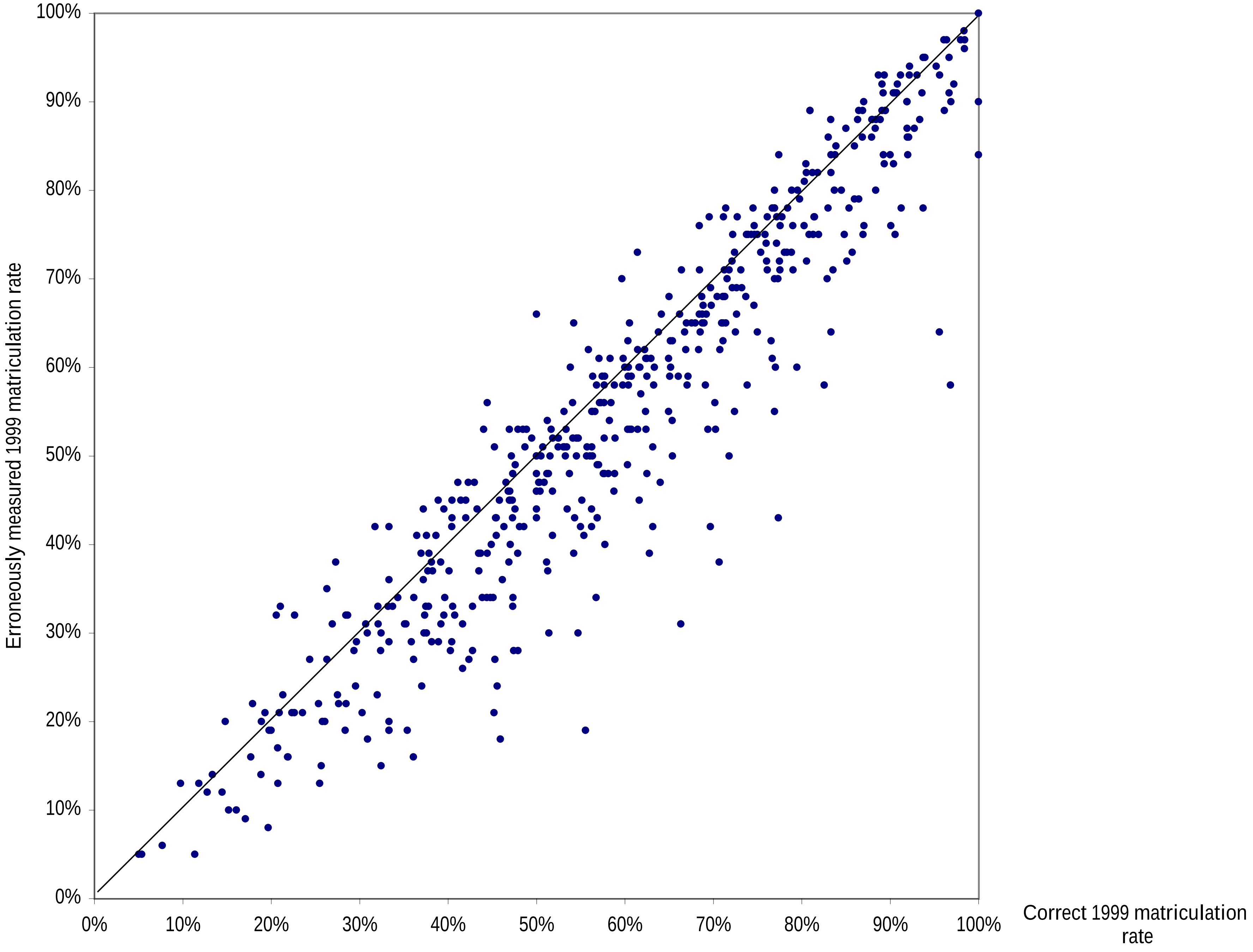
Sadowski, Michael and Miller, Edward. (1996). "New Ideas Like Collective Incentives and Skill-Based Pay Raise the Same Old Questions." The Harvard Education Letter. January/February, 1996.

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.

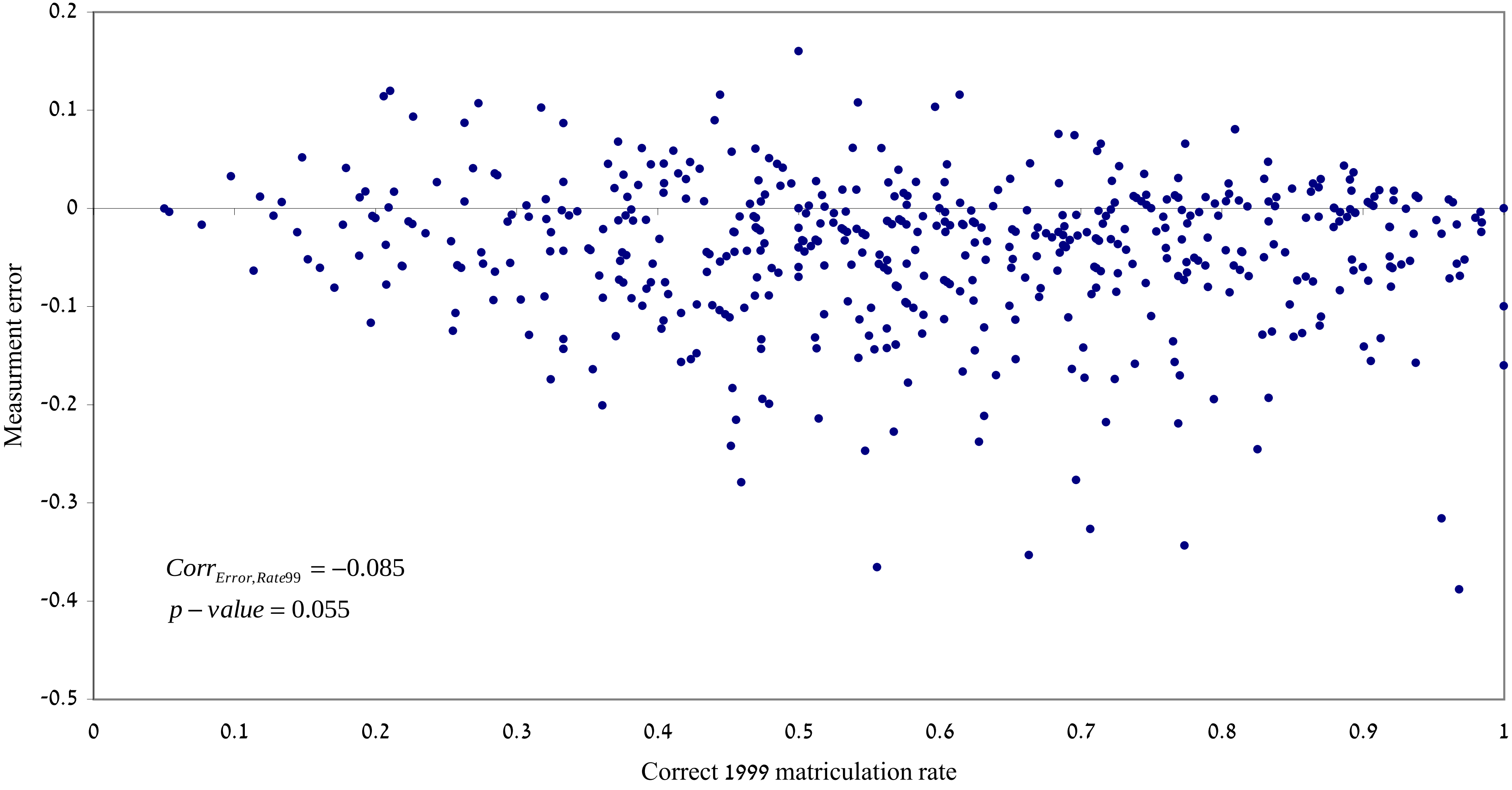
Wakelyn, David J. (1996). The Politics of Compensation Reform: A Colorado Case Study. American Educational Finance Association conference paper, March 23, 1996.

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

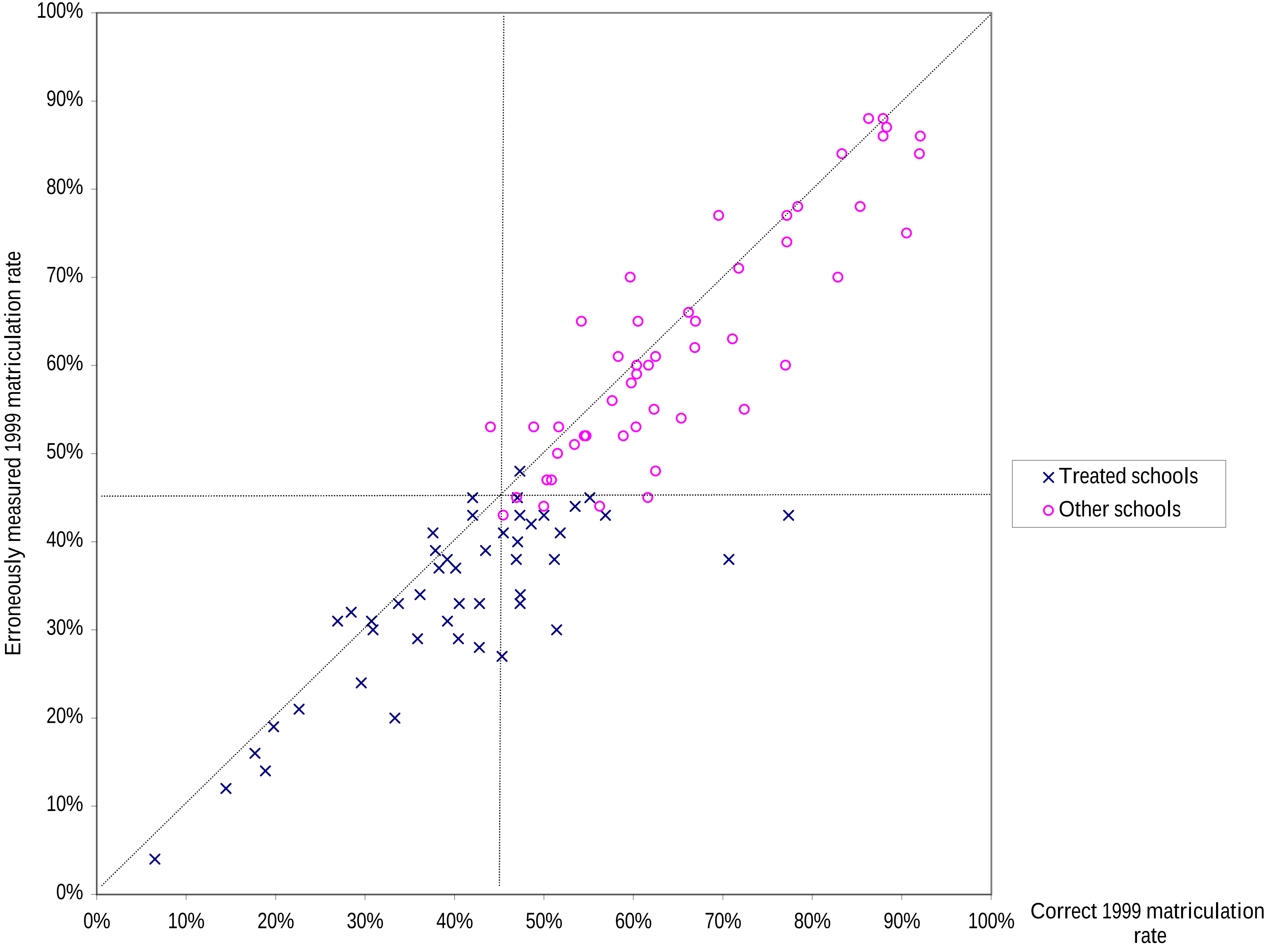


## Figure 2: The Correct 1999 Matriculation Rate Versus The Measurment Error
### Sample=507 Schools



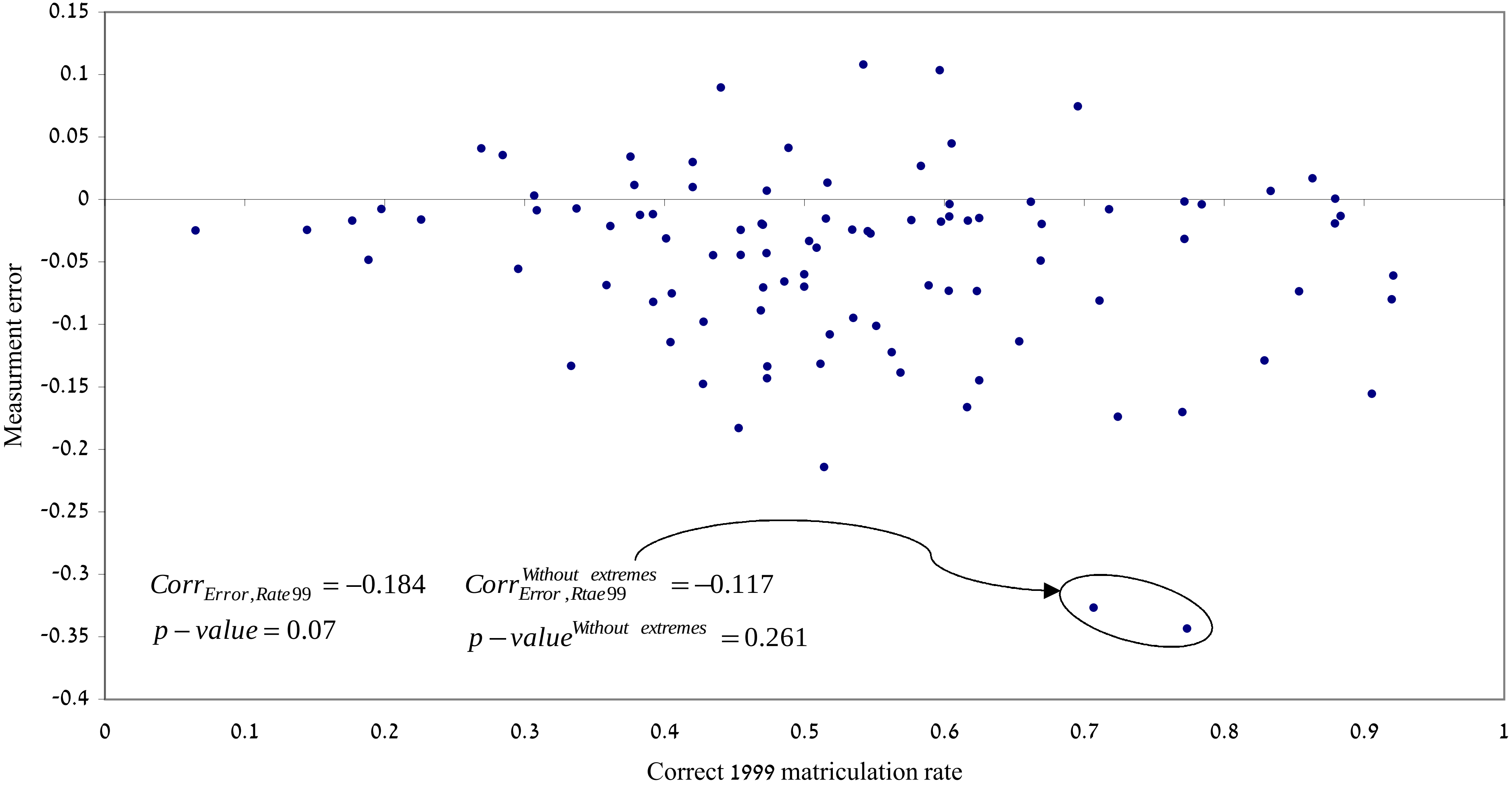
## Figure 3: The Relationship Between the Correct and the Erroneously Measured 1999 Matriculation Rate

Sample=97 Schools



## Figure 4: The Correct 1999 Matriculation Rate Versus The Measurment Error

**Sample=97 Schools**



## Figure 4A: The Correct 1999 Matriculation Rate Versus The Measurment Error

Sample=69 Schools

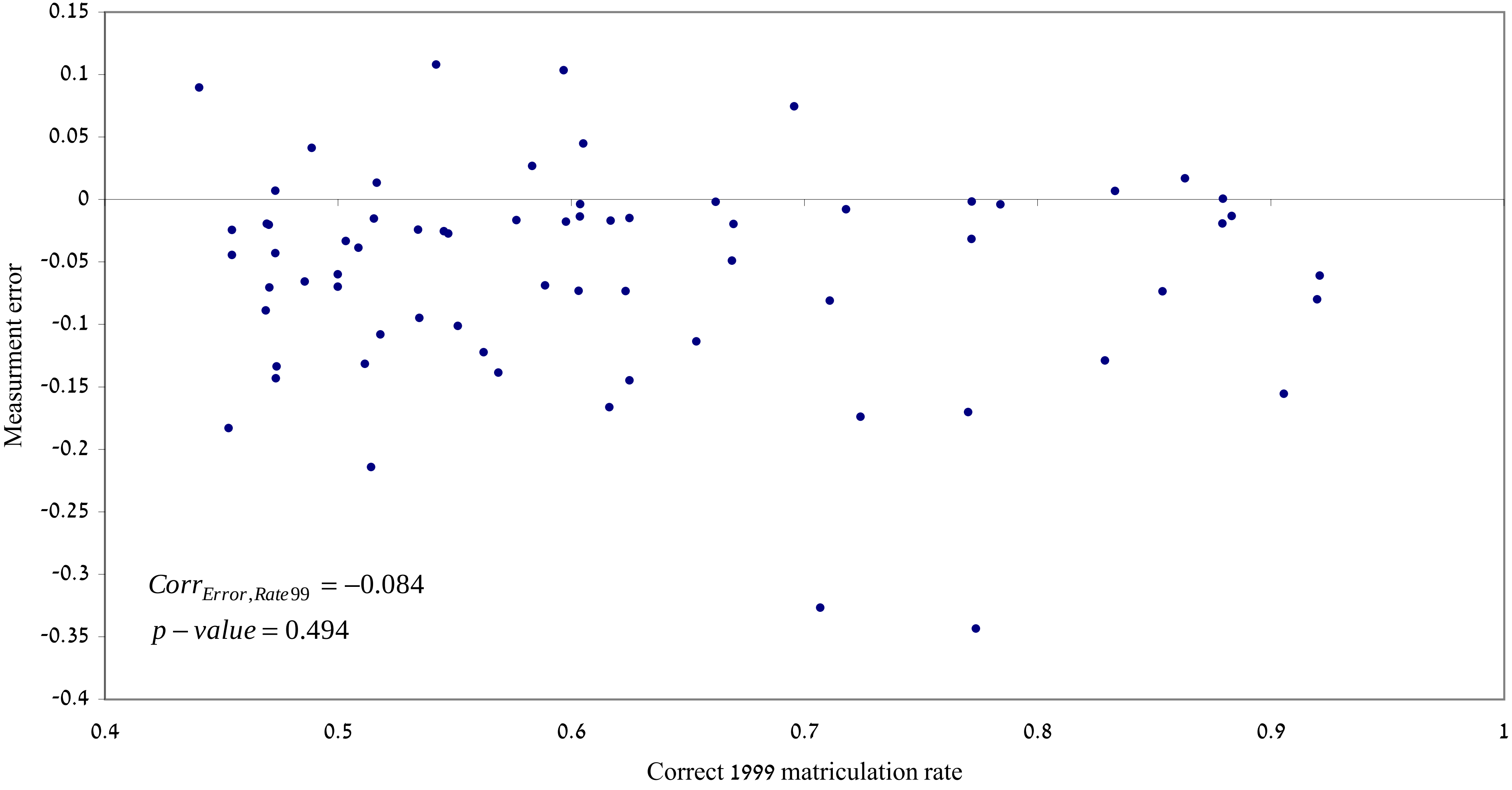


**Figure 5: Determining the Sample of Schools That Were Randomly Assigned To Treatment or Control**

**Sample=97 Schools**

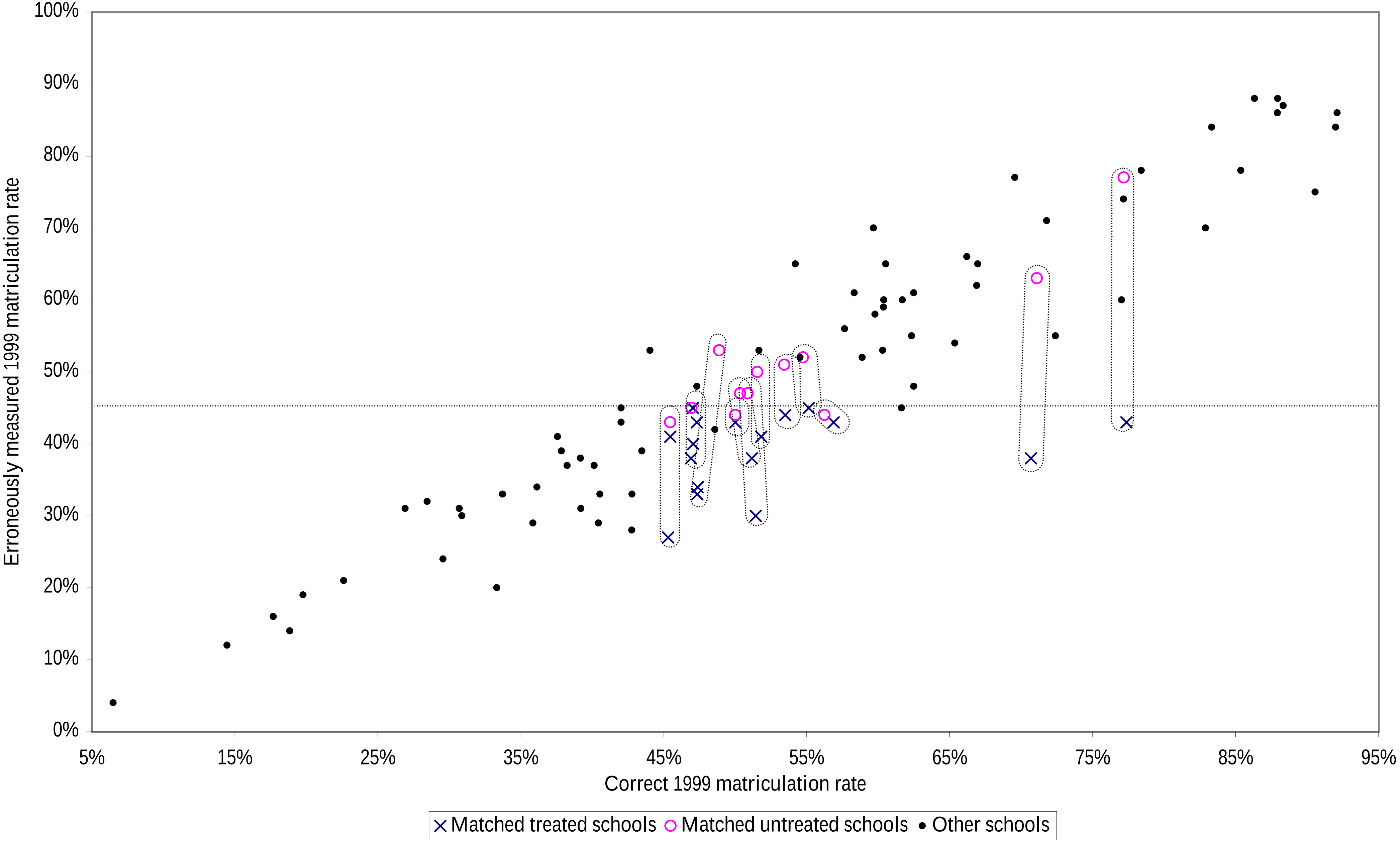


**Figure 6: Determining the Discontinuity Sample (Schools Close To the Threshold Value)**
**Sample=97 Schools**

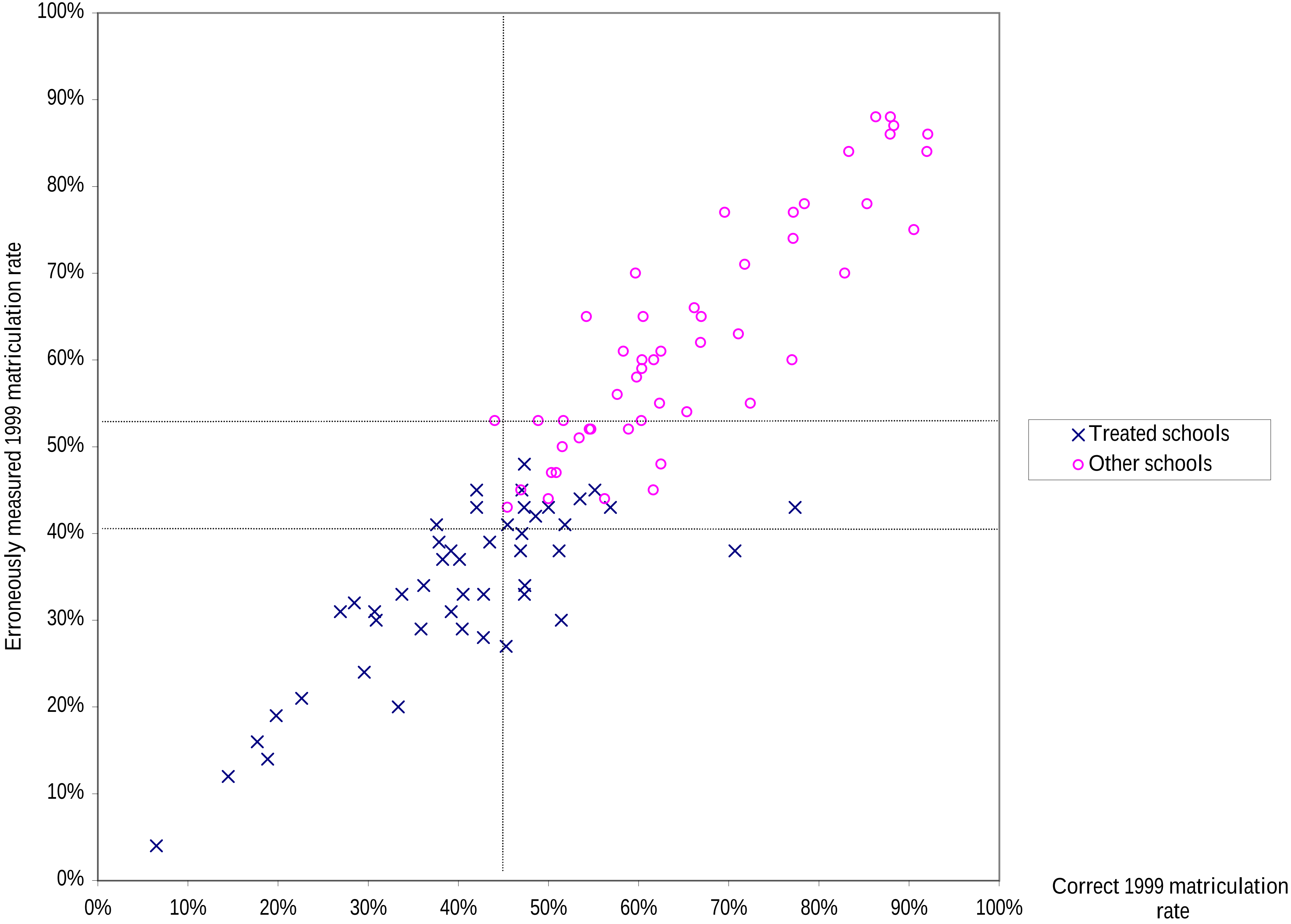


Table 1

Estimated Correlations Between the 1999 Measurement Error 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 | 0.000 | 0.000 |
| | (0.001) | (0.002) | (0.001) | (0.003) |
| Mother's education | 0.000 | -0.001 | -0.001 | -0.003 |
| | (0.001) | (0.002) | (0.001) | (0.002) |
| Number of sibblings | 0.003 | 0.005 | 0.001 | 0.004 |
| | (0.002) | (0.004) | (0.002) | (0.004) |
| Gender (Male=1) | -0.003 | -0.021 | -0.005 | -0.016 |
| | (0.013) | (0.033) | (0.013) | (0.033) |
| Immigrant | -0.035 | -0.132 | -0.002 | -0.459 |
| | (0.036) | (0.093) | (0.045) | (0.164) |
| **Lagged student's outcomes** | | | | |
| Math credits | -0.002 | -0.011 | 0.000 | 0.001 |
| | (0.004) | (0.013) | (0.004) | (0.013) |
| English credits | 0.006 | 0.066 | 0.013 | 0.089 |
| | (0.009) | (0.037) | (0.008) | (0.040) |
| History credits | 0.004 | 0.050 | 0.004 | 0.028 |
| | (0.009) | (0.027) | (0.007) | (0.017) |
| Biology credits | -0.085 | 0.204 | -0.116 | 0.148 |
| | (0.059) | (0.204) | (0.064) | (0.152) |
| Total credits | 0.000 | -0.004 | 0.001 | 0.002 |
| | (0.002) | (0.004) | (0.002) | (0.004) |
| Average score | 0.000 | 0.000 | 0.0006 | 0.001 |
| | (0.000) | (0.001) | (0.0003) | (0.001) |
| **School characteristics** | | | | |
| Religious schools | -0.006 | -0.025 | -0.006 | -0.025 |
| | (0.007) | (0.016) | (0.007) | (0.016) |
| Arab school | 0.025 | 0.024 | 0.025 | 0.024 |
| | (0.009) | (0.019) | (0.009) | (0.019) |
| Number of schools | 507 | 97 | 507 | 97 |

Note: The coefficents 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.

Table 2

Descriptive Statistics: The Measurement Error 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<br>(1.007) | 10.188 | 11.054 | -0.865<br>(0.757) |
| Mother's education | 10.315 | 10.340 | -0.024<br>(1.061) | 10.181 | 10.280 | -0.099<br>(1.082) |
| Number of sibblings | 3.058 | 2.406 | 0.653<br>(0.351) | 3.053 | 1.993 | 1.061<br>(0.389) |
| Gender (Male=1) | 0.494 | 0.505 | -0.011<br>(0.058) | 0.534 | 0.517 | 0.018<br>(0.057) |
| Immigrant | 0.017 | 0.029 | -0.011<br>(0.027) | 0.015 | 0.014 | 0.001<br>(0.014) |
| Asia-Africa ethnicity | 0.228 | 0.287 | -0.059<br>(0.057) | 0.216 | 0.260 | -0.045<br>(0.047) |
| **Lagged student's outcomes** | | | | | | |
| Math credits | 0.375 | 0.557 | -0.182<br>(0.178) | 0.320 | 0.583 | -0.264<br>(0.153) |
| English credits | 0.175 | 0.148 | 0.026<br>(0.060) | 0.138 | 0.123 | 0.015<br>(0.083) |
| History credits | 0.131 | 0.403 | -0.271<br>(0.084) | 0.353 | 0.775 | -0.422<br>(0.161) |
| Biology credits | 0.000 | 0.000 | 0.000<br>(0.000) | 0.000 | 0.000 | 0.000<br>(0.000) |
| Total credits | 4.055 | 4.256 | -0.201<br>(0.443) | 4.111 | 4.420 | -0.309<br>(0.413) |
| **School characteristics** | | | | | | |
| Religious school | 0.296 | 0.205 | 0.091<br>(0.158) | 0.307 | 0.206 | 0.102<br>(0.160) |
| Arab school | 0.165 | 0.000 | 0.165<br>(0.098) | 0.164 | 0.000 | 0.164<br>(0.100) |
| Previous year Bagrut rate (1999,2000) | 0.501 | 0.552 | -0.051<br>(0.032) | 0.479 | 0.498 | -0.019<br>(0.041) |
| Number of observations | 2405 | 1773 | | 2350 | 1678 | |

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 3

The Treatment Effect on English and math Bagrut Outcomes Estimated Using the Natural Experiment (the "Measurement Error" Sample)

| | **Attempted exams** | **Attempted credits** | **Awarded credits** | **Attempted exams** | **Attempted credits** | **Awarded credits** |
|---|---|---|---|---|---|---|
| **Treated Subjects** | **Math** | | | **English** | | |
| Sample mean | 1.20 | 2.09 | 1.61 | 0.94 | 2.51 | 2.06 |
| Control for correct | 0.078 | 0.135 | 0.256 | 0.027 | 0.224 | 0.361 |
| matriculation rate | (0.030) | (0.058) | (0.076) | (0.038) | (0.103) | (0.111) |
| No control for correct | 0.079 | 0.125 | 0.163 | 0.035 | 0.194 | 0.327 |
| matriculation rate | (0.032) | (0.051) | (0.068) | (0.031) | (0.087) | (0.104) |
| **Untreated Subjects** | **History** | | | **Biology** | | |
| Sample mean | 0.33 | 0.35 | 0.22 | 0.68 | 0.08 | 0.06 |
| Control for correct | -0.084 | -0.066 | -0.077 | 0.111 | 0.251 | 0.136 |
| matriculation rate | (0.041) | (0.040) | (0.042) | (0.102) | (0.135) | (0.088) |
| No control for correct | -0.081 | -0.064 | -0.087 | 0.173 | 0.160 | 0.093 |
| matriculation rate | (0.048) | (0.047) | (0.050) | (0.087) | (0.121) | (0.076) |

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. Students level controls include the number of sibblings, gender dummy, father's and other'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 nits gained before treatment and the average score in the relevant tests. School fixed effects are included as well in each of the regressions.

Table 3a

Effects on Englsih and math Bagrut Outcomes by Quartiles of Previous Test Scores, Using the 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 | 0.184 | 0.017 | -0.036 | 0.093 | -0.009 | -0.053 | -0.009 |
| | (0.083) | (0.056) | (0.060) | (0.067) | (0.095) | (0.058) | (0.073) | (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 | 0.365 | 0.063 | -0.141 | 0.497 | 0.357 | 0.020 | -0.127 |
| | (0.123) | (0.105) | (0.097) | (0.123) | (0.186) | (0.141) | (0.160) | (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 | 0.499 | 0.334 | -0.011 | 0.707 | 0.581 | 0.095 | -0.084 |
| | (0.114) | (0.103) | (0.102) | (0.114) | (0.160) | (0.151) | (0.153) | (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 | 0.076 | 0.013 | -0.064 | * | * | * | * |
| | (0.028) | (0.038) | (0.031) | (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). 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 4

Descriptive Statistics: The Discontinuity Sample

| | Year 2000 | | | Year 2001 | | |
|---|---|---|---|---|---|---|
| | Treatment | Control | Difference<br>(s.e) | Treatment | Control | Difference<br>(s.e) |
| **Student's background** | | | | | | |
| Father's education | 11.055 | 10.424 | 0.631 | 10.889 | 10.455 | 0.434 |
| | | | (0.486) | | | (0.511) |
| Mother's education | 11.124 | 10.733 | 0.391 | 11.088 | 10.882 | 0.206 |
| | | | (0.561) | | | (0.575) |
| Number of sibblings | 2.609 | 2.427 | 0.182 | 2.552 | 2.131 | 0.421 |
| | | | (0.344) | | | (0.393) |
| Gender (Male=1) | 0.492 | 0.468 | 0.024 | 0.498 | 0.488 | 0.010 |
| | | | (0.066) | | | (0.062) |
| Immigrant | 0.013 | 0.045 | -0.032 | 0.013 | 0.008 | 0.005 |
| | | | (0.022) | | | (0.007) |
| Asia-Africa ethnicity | 0.218 | 0.319 | -0.101 | 0.211 | 0.287 | -0.077 |
| | | | (0.050) | | | (0.054) |
| **Lagged student's outcomes** | | | | | | |
| Math credits | 0.229 | 0.568 | -0.339 | 0.265 | 0.578 | -0.312 |
| | | | (0.147) | | | (0.176) |
| English credits | 0.208 | 0.088 | 0.120 | 0.185 | 0.125 | 0.060 |
| | | | (0.061) | | | (0.090) |
| History credits | 0.155 | 0.176 | -0.022 | 0.434 | 0.567 | -0.133 |
| | | | (0.084) | | | (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 | 4.230 | 4.594 | -0.364 |
| | | | (0.346) | | | (0.388) |
| **School characteristics** | | | | | | |
| Religious school | 0.098 | 0.325 | -0.227 | 0.092 | 0.309 | -0.217 |
| | | | (0.152) | | | (0.150) |
| Arab school | 0.128 | 0.000 | 0.128 | 0.128 | 0.000 | 0.128 |
| | | | (0.090) | | | (0.091) |
| Previous year Bagrut rate | 0.483 | 0.537 | -0.054 | 0.482 | 0.507 | -0.025 |
| (1999,2000) | | | (0.016) | | | (0.042) |
| 1999 measured with error | 0.426 | 0.488 | -0.061 | - | - | - |
| Bagrut rate | | | (0.011) | | | |
| Number of observations | 2523 | 1564 | | 2535 | 1406 | |

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 5

The Treatment Effect Estimated Using the Discontinuity Sample

| | **Attempted exams** | **Attempted credits** | **Awarded credits** | **Attempted exams** | **Attempted credits** | **Awarded credits** |
|---|---|---|---|---|---|---|
| **Treated Subjects** | | **Math** | | | **English** | |
| Sample mean | 1.22 | 2.14 | 1.74 | 0.89 | 2.53 | 2.16 |
| Control for correct | 0.047 | 0.100 | 0.244 | 0.129 | 0.212 | 0.177 |
| matriculation rate | (0.030) | (0.056) | (0.078) | (0.031) | (0.081) | (0.104) |
| No control for correct | 0.046 | 0.093 | 0.231 | 0.132 | 0.199 | 0.177 |
| matriculation rate | (0.032) | (0.058) | (0.079) | (0.030) | (0.079) | (0.108) |
| **Untreated Subjects** | | **History** | | | **Biology** | |
| Sample mean | 0.51 | 0.53 | 0.22 | 0.58 | 0.05 | 0.05 |
| Control for correct | 0.138 | 0.160 | 0.064 | 0.146 | -0.114 | -0.140 |
| matriculation rate | (0.084) | (0.083) | (0.079) | (0.089) | (0.146) | (0.075) |
| No control for correct | 0.122 | 0.144 | 0.037 | 0.147 | -0.127 | -0.152 |
| matriculation rate | (0.086) | (0.089) | (0.088) | (0.088) | (0.141) | (0.071) |

Note: The table reports treatment-control differences for thr 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. Students level controls include the number of sibblings, gender dummy, father's and other'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 5a

Effects on English and Math Bagrut Outcomes by Quartiles of Previous Test Scores, 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 | 0.192 | 0.144 | 0.095 | 0.156 | 0.166 | 0.060 | 0.132 |
| | (0.075) | (0.059) | (0.052) | (0.064) | (0.088) | (0.056) | (0.057) | (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 | 0.246 | 0.125 | -0.068 | 0.313 | 0.414 | -0.016 | 0.132 |
| | (0.124) | (0.114) | (0.087) | (0.118) | (0.171) | (0.089) | (0.130) | (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 | 0.361 | 0.391 | 0.060 | 0.375 | 0.435 | -0.100 | -0.030 |
| | (0.118) | (0.114) | (0.093) | (0.111) | (0.161) | (0.132) | (0.137) | (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 | 0.089 | 0.036 | -0.020 | * | * | * | * |
| | (0.023) | (0.030) | (0.026) | (0.034) | * | * | * | * |
| Control group mean | 0.063 | 0.380 | 0.690 | 0.831 | * | * | * | * |

Note: The table reports treatment 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). 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 6

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 | 9.471 | 9.296 | 0.175 |
| | | | (0.550) | | | (0.541) |
| Mother's education | 8.831 | 8.733 | 0.097 | 9.149 | 8.918 | 0.231 |
| | | | (0.697) | | | (0.673) |
| Number of sibblings | 3.342 | 3.483 | -0.140 | 3.217 | 3.296 | -0.079 |
| | | | (0.394) | | | (0.387) |
| Gender (Male=1) | 0.488 | 0.471 | 0.017 | 0.483 | 0.471 | 0.012 |
| | | | (0.025) | | | (0.024) |
| Immigrant | 0.014 | 0.001 | 0.014 | 0.014 | 0.001 | 0.014 |
| | | | (0.004) | | | (0.004) |
| Asia-Africa ethnicity | 0.165 | 0.185 | -0.021 | 0.179 | 0.210 | -0.031 |
| | | | (0.028) | | | (0.027) |
| **Lagged student's outcomes** | | | | | | |
| Math credits | 0.362 | 0.304 | 0.058 | 0.531 | 0.535 | -0.004 |
| | | | (0.081) | | | (0.109) |
| English credits | 0.109 | 0.160 | -0.051 | 0.060 | 0.051 | 0.009 |
| | | | (0.047) | | | (0.023) |
| History credits | 0.442 | 0.409 | 0.033 | 0.453 | 0.419 | 0.034 |
| | | | (0.069) | | | (0.069) |
| Biology credits | 0.000 | 0.002 | -0.002 | 0.000 | 0.002 | -0.002 |
| | | | (0.001) | | | (0.001) |
| Total credits | 4.158 | 4.101 | 0.057 | 4.311 | 4.256 | 0.055 |
| | | | (0.288) | | | (0.259) |
| Average Score | 64.225 | 63.416 | 0.809 | 65.047 | 63.559 | 1.488 |
| | | | (2.045) | | | (1.857) |
| **School characteristics** | | | | | | |
| Religious school | 0.179 | 0.203 | -0.024 | 0.171 | 0.185 | -0.015 |
| | | | (0.064) | | | (0.062) |
| Arab school | 0.286 | 0.326 | -0.041 | 0.232 | 0.253 | -0.020 |
| | | | (0.090) | | | (0.081) |
| Previous year Bagrut rate (2000) | 0.409 | 0.423 | -0.014 | 0.422 | 0.426 | -0.004 |
| | | | (0.024) | | | (0.023) |
| Number of observations | 4490 | 4490 | | 4865 | 4865 | |

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 6a

Descriptive statistics: the Propencity score sample based on all students in participating schools and their matches.

| | All matching | | |
|---|---|---|---|
| | treatment | control | Difference (s.e) |
| **Student's background** | | | |
| Father's education | 9.362 | 9.248 | 0.114<br>(0.528) |
| Mother's education | 9.005 | 8.774 | 0.232<br>(0.658) |
| Number of sibblings | 3.220 | 3.342 | -0.122<br>(0.363) |
| Gender (Male=1) | 0.483 | 0.468 | 0.015<br>(0.022) |
| Immigrant | 0.013 | 0.001 | 0.013<br>(0.004) |
| Asia-Africa ethnicity | 0.182 | 0.215 | -0.033<br>(0.027) |
| **Lagged student's outcomes** | | | |
| Math credits | 0.498 | 0.448 | 0.050<br>(0.100) |
| English credits | 0.101 | 0.090 | 0.012<br>(0.038) |
| History credits | 0.442 | 0.431 | 0.011<br>(0.066) |
| Biology credits | 0.000 | 0.000 | 0.000<br>(0.000) |
| Total credits | 4.196 | 4.256 | -0.060<br>(0.262) |
| Average Score | 63.921 | 64.132 | -0.210<br>(1.797) |
| **School characteristics** | | | |
| Religious school | 0.178 | 0.199 | -0.021<br>(0.061) |
| Arab school | 0.247 | 0.284 | -0.037<br>(0.081) |
| Previous year Bagrut rate (2000) | 0.417 | 0.425 | -0.009<br>(0.023) |
| Number of observations | 5512 | 5512 | |

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

Table 7

The Treatment Effect Estimated By the Propensity Sample

| | Attempted exams | Attempted credits | Awarded credits | Attempted exams | Attempted credits | Awarded credits |
|---|---|---|---|---|---|---|
| **Treated Subjects** | **Math** | | | **English** | | |
| **Treated students** | | | | | | |
| Sample mean | 1.34 | 2.33 | 1.82 | 1.08 | 2.78 | 2.24 |
| | 0.242 | 0.398 | 0.293 | 0.088 | 0.230 | 0.145 |
| | (0.032) | (0.059) | (0.068) | (0.033) | (0.068) | (0.079) |
| All students | | | | | | |
| Sample mean | 1.19 | 2.07 | 1.60 | 1.05 | 2.66 | 2.13 |
| | 0.149 | 0.230 | 0.151 | 0.091 | 0.177 | 0.120 |
| | (0.039) | (0.063) | (0.065) | (0.033) | (0.070) | (0.076) |
| **Untreated Subjects** | **History** | | | **Biology** | | |
| **Treated students** | | | | | | |
| Sample mean | 0.51 | 0.52 | 0.25 | 0.72 | 0.11 | 0.05 |
| | 0.158 | 0.145 | 0.020 | 0.156 | 0.062 | 0.024 |
| | (0.084) | (0.084) | (0.046) | (0.090) | (0.034) | (0.019) |
| All students | | | | | | |
| Sample mean | 0.50 | 0.51 | 0.26 | 0.73 | 0.11 | 0.05 |
| | 0.145 | 0.125 | 0.028 | 0.192 | 0.060 | 0.030 |
| | (0.080) | (0.080) | (0.045) | (0.081) | (0.032) | (0.019) |

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. Students level controls include the number of sibblings, gender dummy, father's and other'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 7a

Effects on Bagrut Englsih and Math Outcomes by Quartiles of Previous Test Scores, Propensity Score Sample Based on All Students in Participating Schools and Their Matches

| | 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 | 0.289 | 0.129 | 0.079 | 0.104 | 0.249 | 0.010 | -0.025 | 0.081 |
| | (0.064) | (0.050) | (0.035) | (0.049) | (0.060) | (0.047) | (0.043) | (0.050) |
| Control group mean | 0.544 | 1.139 | 1.233 | 1.341 | 0.813 | 1.192 | 1.042 | 0.867 |
| **Attempted credits** | | | | | | | | |
| Treatment effect | 0.450 | 0.240 | 0.177 | 0.126 | 0.391 | 0.110 | 0.038 | 0.171 |
| | (0.100) | (0.082) | (0.071) | (0.105) | (0.108) | (0.087) | (0.090) | (0.124) |
| Control group mean | 0.823 | 1.862 | 2.162 | 2.626 | 1.506 | 2.643 | 2.834 | 2.884 |
| **Awarded credits** | | | | | | | | |
| Treatment effect | 0.262 | 0.243 | 0.172 | 0.048 | 0.230 | 0.161 | 0.007 | 0.124 |
| | (0.075) | (0.088) | (0.083) | (0.106) | (0.096) | (0.099) | (0.101) | (0.136) |
| Control group mean | 0.420 | 1.299 | 1.773 | 2.414 | 0.911 | 1.994 | 2.406 | 2.581 |
| **Bagrut rate** | | | | | | | | |
| Treatment effect | 0.023 | 0.075 | 0.016 | 0.042 | * | * | * | * |
| | (0.018) | (0.025) | (0.024) | (0.032) | * | * | * | * |
| Control group mean | 0.055 | 0.374 | 0.652 | 0.774 | * | * | * | * |

Note: The table reports treatment 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 ath 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.

* 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 8

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

| | English teachers | | Math teachers | |
|---|---|---|---|---|
| | Sample mean | Treatment-control difference | Sample mean | Treatment-control difference |
| **Teaching methods:** | | | | |
| Teaching in small groups | 0.668 | 0.085<br>(0.052) | 0.665 | 0.007<br>(0.050) |
| Individualized instruction | 0.631 | 0.112*<br>(0.054) | 0.601 | -0.028<br>(0.052) |
| Tracking by ability | 0.512 | 0.221*<br>(0.055) | 0.458 | 0.13*<br>(0.052) |
| Adapting teaching methods to students ability | 0.954 | 0.068*<br>(0.023) | 0.947 | 0.011<br>(0.024) |
| **Teacher's effort:** | | | | |
| Added instruction time during the year | 0.662 | 0.220*<br>(0.052) | 0.838 | 0.015*<br>(0.039) |
| Added instruction time before Bagrut exam | 0.412 | 0.146*<br>(0.067) | 0.390 | 0.070<br>(0.056) |
| Number of weeks before Bagrut exam with added instruction time | 4.808 | 0.564<br>(0.877) | 5.176 | 2.159*<br>(0.964) |
| Number of additional instruction hours | 3.068 | 0.829<br>(0.657) | 3.918 | 2.102*<br>(0.769) |
| **Teacher's effort targeted towards:** | | | | |
| All students | 0.328 | -0.004<br>(0.052) | 0.575 | -0.025<br>(0.052) |
| Weak students | 0.252 | 0.129*<br>(0.048) | 0.184 | -0.058<br>(0.041) |
| Average students | 0.024 | 0.019<br>(0.017) | 0.031 | 0.043*<br>(0.018) |
| Strong students | 0.012 | 0.028*<br>(0.012) | 0.003 | 0.006<br>(0.006) |
| 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.

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

Note: Standard errors in parenthesis. Asteriskes 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.

Elite universities: Hebrew University in Jerusalem, Tel-Aviv, Technion and Weizman Institute. Other Universities: Bar-Ilan, Ben Gurion and Haifa university.