Evaluating the Effect of Teachers' Performance Incentives on Pupils' Achievements

Victor Lavy
Hebrew University of Jerusalem

June, 1999

ABSTRACT

Proposals to use teachers' performance incentives as the basis for school reforms have recently attracted considerable attention and support among researchers and policy makers. The main message is that the most likely way to improve students' achievements is to institute performance incentives, direct monetary rewards for improvements in student outcomes. However, there has been very little experience with applying performance incentives in schools. This paper provides empirical evidence on the causal effect of a program that offered monetary incentives to teachers as a function of their students' achievements. The program offered incentives to schools in the form of performance awards, part of which were distributed to teachers and school staff as merit pay and the rest, for the well-being of teachers in the school. I evaluate the effect of the program during the first two full academic years of its implementation, 1996 and 1997. Since participating schools were not chosen randomly, the issue of identification is central to the empirical strategy. The second part of the paper evaluates the effect of a 'twin' program, implemented simultaneously and with the same objectives as the incentive program, but in different schools, and based on providing additional resources to schools. The paper compares the effect and cost of the two alternative intervention strategies and draws policy implications.

Victor Lavy
Department of Economics, Hebrew University
Mt. Scopus, Jerusalem 91905, Israel
msvictor@mscc.huji.ac.il

* I owe special thanks to the staff at the Ministry of Education for providing and assisting with the data used in this study, and to J. Angrist, D. Genesove, C. Manski and S. Pischke for helpful comments and discussions. I also benefited from comments and suggestions of seminar participants at Harvard, MIT, The Hebrew University and The World Bank. Special thanks to Analia Schlosser for her excellent research assistance. The author bears sole responsibility for the contents of this paper. It does not reflect the views of the Government of Israel or those of any of its agencies or employees.

I. Introduction

Proposals to use teachers' or school performance incentives as the basis for school reforms have recently attracted considerable attention and support among researchers and policy makers.[1] The main message in the relevant body of literature is that the most likely promising way to improve students' achievements is to institute monetary performance incentives, direct rewards for improvements in student outcomes. It is argued that alternative interventions, implemented in the past, tend to increase costs with no commensurate improvement in student performance. However, there has been very little experience with applying performance incentives to schools.[2] Therefore, many authors emphasize that before making the introduction of school incentives the next revolution in schools, much more concrete evidence is needed about the optimal incentive structure in schools and their effect and cost.

This paper seeks to derive empirical evidence on the causal effect of a program that introduced monetary incentives to teachers as a function of the relative improvement in their students' achievements.

---

[1] Two recently published books focus on the role of incentives in improving American schools: Hanushek *et al.* (1994) and Hanushek and Jorgenson (1996). The articles by Hanushek, Hannaway, and Bishop in the latter volume directly address theoretical aspects of various schemes of incentives and student performance. See also Heckman (1998).

[2] Various teacher- and school-incentive programs were implemented in the U.S., some as State initiatives (Kentucky, South Carolina) and others as local initiatives (Dallas, Baltimore). Details on these and other programs are presented in Clotfeller and Ladd (1996) and Elmore, Abelmann and Fuhrman (1996). Some of these programs, especially those based on teachers' merit pay, have been studied (see a review by Cohen and Murnane, 1985), and the results are not encouraging, being either inconclusive or suggesting no improvement. Incentives to teachers and schools in the form of performance awards have been few. Richards and Sheu (1992), for example, analyze the school incentives-reward program in South Carolina. The authors find modest improvements in students' achievements and no significant improvement in attendance patterns of students or teachers. Note, however, that the studies on merit pay and performance incentives lack the strength of experimental evidence and none of them provide convincing estimates of the causal effect of these programs. While studies in economics suggest that incentives in private firms have a significant effect on effort and output (Lazear, 1996), these programs are seldom evaluated systematically (Milgrom and Roberts, 1992).

The program was implemented in 62 secondary schools in Israel, which were selected non-randomly to the program in 1995. The program offered incentives to schools in the form of performance awards, part of which were distributed to teachers and school staff as merit pay, and the rest used for the well-being (upgrading general work conditions) of teachers in the school. The total sum awarded was determined in advance (about $1.4m) and was distributed among the top third performers. The competition was based on various scholastic achievement criteria, including dropout rates.

This teachers' incentives program fits closely the framework of a rank order tournament analyzed in Lazear and Rosen (1981) and Green an Stokey (1983): the scheme pays prizes to the winners of a contest or a tournament in which the prizes depends on the rank order of the winner. In this particular contest the distance between winners is also taken into account in determining how the fix prize is distributed among winners. Comparison of the results in this paper with some relevant predictions from the theory may be insightful.

This paper presents an evaluation of the effect of the program during the first two full academic years of its implementation, 1996 and 1997. Since the participating schools were not chosen at random, the issue of identification is central to the empirical strategy. Indeed, participating ('treated') schools differed considerably, in terms of pre-program characteristics and outcomes, from all other secondary schools in Israel. Nonetheless, the Ministry of Education's rules of selection into the program provide a potential quasi-natural experiment that could form the basis for credible identification strategy. Israel's Hebrew secondary state school system includes separate religious girls' and boys' schools and religious and non-religious coeducational schools. The program was targeted at all schools that were the only one of their kind in each local area (community). For example, if a particular religious girls' school was selected, there was no other such school in that community. The number of alternatives of a given school type in a community varies in Israel from 1 to over 25. Selection into the program can therefore be described as a

discontinuous (threshold) function of this observed distribution. The schools around the cut-off point, those included in the group of 'one' and 'two' schools in the community, might be very similar in outcomes and characteristics and therefore may form the basis for identification. This identification strategy can be viewed as a natural experiment or as a natural application of the Campbell (1969) discontinuity-design method.[3]

As noted above, the 'incentive' approach to school reform is suggested as an alternative to the 'resources' approach, whereby resources are added to schools and specific interventions are structured that hopefully will lead to better outcomes. A comparison of the two approaches in terms of their effect and cost is essential in order to formulate policy implications. The second part of this paper therefore evaluates the effect of a 'twin' program based on the 'resources' approach and compares its effect and cost to the incentive-based program. Alongside the teachers' incentives program, Israel's Ministry of Education conducted a three-year programthat endowed 22 high schools with additional resources, mainly teaching time and on-the-job school staff training. Here, too, the objective was to improve students' performance, and the 22 participating schools (selected from 75 applicants) planned and developed the intervention individually and independently. I evaluate the effect of this parallel program in its three years of implementation, and compare its effectiveness and cost with its 'twin', the teachers' incentives intervention. This comparison has the attractive feature that the two programs apply at the school level: all teachers in the school face the same incentives and the resources are used at the school level. How is it made to work within schools is unknown, for both programs.

The identification strategy is based on a comparison between treated schools and other applicants,

[3] Van der Klaauw (1996), Angrist and Lavy (1999) and Lavy (1998) are recent examples of the application of the regression discontinuity method in economics.

which were not admitted to the program.[4] Treated schools differed considerably in pre-program characteristics and outcomes from other Israeli schools, but were similar to other unselected applicants. To control for remaining observed and unobserved permanent differences between the treated and comparison schools, panel data on schools are used to estimate school fixed effect models.

The data used in the study include characteristics and scholastic achievements of all high-school graduates in Israel in 1993-97. The data are a panel at the school level. The teachers' monetary incentives are a function of the achievements of students in their final year of high school and of the dropout rate at all high-school grades. Panel data on students were used in order to determine the dropout rate at each grade.

The paper is organized as follows. Following a description of the two programs and the Israeli high school scholastic achievement data (Section II), Section III presents a simple graphical presentation of the identification strategy of the incentive program. Section IV describes the statistical model used for inference for both programs. Section V presents the main results for the two intervention strategies, and Section VI presents a cost-benefit comparison of the two policy alternatives. Section VII concludes. The results suggest that teachers' monetary incentives had some effect in the first year of implementation (mainly in religious schools), it caused significant gains in many dimensions of students' outcomes in the second year (in religious and non-religious schools alike). However, endowing schools with more conventional resources, such as additional teaching time and on-the-job teacher training, also led to significant improvement in student performance. The comparison based on cost-equivalency suggests that the teachers' incentive intervention is much more cost effective.

II. The Interventions and the Data

The two intervention strategies were designed to improve the performance of high-school graduates in their

[4] There is no overlap between the four groups – the two treatment groups and the two contorl groups.

matriculation exams and to reduce dropout rates at different grades in high school.[5] The post-primary state school system in Israel has two types of secondary school: the first, called 'secondary comprehensive' schools, includes both the middle- and high school grades (grades 7-12); the second includes only high-school grades (10-12). About 65 percent of all high-school students are enrolled in secondary comprehensive schools. The incentives project was targeted exclusively at these secondary comprehensive schools, while the resources project included both types of school.

Towards the end of the last year of high-school students sit for exams in various subjects, some compulsory and some elective. In the 11th grade students choose one of three difficulty levels of study in each subject, each awarding a different number of credit units: 3, 4 and 5. Choosing level 4 or 5 (credit units) in a given subject earns a bonus of 12.5% or 25%, respectively, in the test score of the matriculation exam in that subject. A minimum of 22 credit units is required to qualify for a matriculation certificate, which is a necessary though not sufficient requirement for admission to university.[6] About 42% of all 1997 high-school graduates received a matriculation certificate.

The teachers' incentive intervention

In February 1995 Israel's Ministry of Education announced a new experimental competition that would reward teachers and schools with monetary bonuses based on their students' performance.[7] The objectives of the program were to reduce dropout rates in secondary schools and improve scholastic achievements. The achievement measures specified were: the average number of credit units per student, the proportion

[5] The matriculation certificate is similar to the Baccalauréate in France and to the "Certificate of Maturity" (*Reifezeugnis*) in Germany.

[6] A weighted average of all the matriculation test scores and the score in a psychometric test serve as admission criteria by universities in Israel.

[7] Details are provided in a publication of the Chief Scientist of the Ministry of Education, The Differential

of students taking the matriculation track,[8] the proportion of students receiving a matriculation certificate, and the school's dropout rate.

Sixty-two schools were selected to the program, with a few more added later. The condition for selection was that the school should be the only one of its kind in its community, and that it be a secondary comprehensive school. In principle, all the Hebrew system schools that met these criteria should have been included in the project, but in practice some eligible schools in the religious Hebrew education system and in the Arab education system were excluded, leaving only the Hebrew secular schools where this rule was strictly implemented. Participating schools competed for a total of about $1.44m in awards in 1996. Schools were ranked each year according to their improvement in the various performance measures, relative to an expected base predicted from regressions controlling for the socio-economic background of the student body. Only the top third performers gained awards. The distribution of cash incentives among award-winning schools was determined according to their ranking in terms of relative improvement (in 1996 the highest scoring school won $105,000 and the lowest award was $13,250). 75% of the award was distributed among the teachers (proportionally to their gross income) as a salary bonus; the remainder was used to improve all-faculty facilities, such as teachers' common rooms. In 1996 the highest bonus per teacher was $715, and the lowest, $200 (the average starting annual salary of a high school teacher is $20,000, the mean is $30,000). Although the program was only announced in the spring of 1995, teachers received awards for the 1994/5 school year, mainly in order to enhance the credibility of the program and draw teachers' attention to the program.

Compensation: Principles for Allocation, 1995, Jerusalem, Israel (Hebrew).

[8] There are two tracks in high school: the matriculation track and the technical track. The latter, offered only in technical schools, leads to a high school completion certificate that is generally deemed inferior to the matriculation certificate in the labor market and in post secondary schooling institutions.

Group performance incentives in economics

Formal economic theory usually justifies incentives to individuals on the grounds that it is the individual who must be motivated to work. In practice, however, the most common explicit incentive contracts are applied across groups of employees. The program evaluated here belongs to the category of group incentives, where the combined performance of a group determines the total incentive payment, which is divided among individuals regardless of individual performance. The most common types of group incentives are profit-sharing and gains-sharing plans. The teachers' incentive plan is a type of gains-sharing plan: when a group meets or exceeds a predetermined target, all its members receive bonuses according to the extent to which goals were exceeded.

There are several reasons why group incentives might be no less (or even more) effective than individual incentives. First, it is often the case that the contribution of an individual member cannot be measured. Secondly, groups of employees often have better information about their constituent individuals and their respective contributions, enabling the group to monitor members and encourage effort or other appropriate behavior. Third, individuals who share 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.[9]

The teachers' performance incentive program considered in this paper does not provide an answer to the free-rider problem. However, since schools are relatively (at least in Israel) small organizations,

[9] Despite the potential free-rider problem, however, group incentive plans are fairly widespread. One explanation is that employees might be reluctant to harm co-workers by shirking on the job. In this case, guilt and social opprobrium might motivate employees even if they do not face any direct financial consequences of shirking (Kandel and Lazear, 1992).

teachers can effectively monitor each other for lack of effort or sub-standard performance.[10] Also, many teachers teach each student, so that cooperation and teamwork are natural in a school. On the other hand, teamwork comes from cohesion, from perfect substitution or coordination among inputs but in schools substitution between teachers is mostly imperfect (for example, the math teacher cannot teach language). The influence of several teachers on each student makes it very difficult to measure individual teacher performance while the performance of a group of teachers can be measured at low cost. In our example, monitoring student performance adds no additional cost since the information is already available as an integral part of the system.

Another complicating factor is the multifaceted nature of teachers' effort in school, which may involve moral hazard issues and strategic behavior on their part (Holmstrom and Milgrom, 1991). Teachers' tasks in the program involve the improvement of several, sometimes-conflicting, performance measures. For example, reducing the dropout rate may result in lower scores in cognitive achievement measures (and qualifying for matriculation certificates) since students who are prevented from dropping out might be less able academically. Another moot element in the program is that not all schools that demonstrate an improvement in performance receive bonuses. This may discourage some schools, especially those with less margin for improvement. For the same reason it may be difficult to achieve high improvement rates in consecutive years.

The 'school resources' program

In July 1994 the Israeli Ministry of Education announced another new experimental project based on endowing schools with additional teaching time and teachers' on-the-job training. This project had the same

[10] Gaynor and Pauly (1990) provide empirical evidence suggesting that group incentives are more effcieint the smaller the group.

objectives as the incentive project. Schools enjoyed complete freedom in designing their intervention; the Ministry intended to identify the successful programs and apply them in other schools. About 70 secondary schools applied to participate in this program. Each applicant submitted a proposal detailing the specific uses to which the additional resources would be put. A Ministerial steering committee reviewed and ranked the applicants, finally selecting 25 schools, but only 22 schools ultimately participated in the program. Hebrew religious and non-religious schools and Arab schools were selected according to their proportion in the population of schools. Schools used the additional resources to add teaching time, to split classes into smaller study groups, to provide extra coaching for weak students and to arrange marathon sessions of intensive preparation for the matriculation exams. Ministry of Education experts provided on-the-job teacher training focused on handling (academic) class heterogeneity. The project started in September 1995 and ended three years later. Each school received a voucher of weekly hours of teachers' time (50 hours for large schools and 30 hours for smaller schools) and about 8 hours a month of teacher training (on the school premises). 50 teaching hours is equivalent to 2.5 additional full time teachers, which is about 3 percent of the mean number of teachers per school in Israel. The total annual cost of the program was about $1.15 million.

The data

The Ministry of Education provided the data for this study. The data are for the three years preceding the two programs, 1993-95, and for the two full years of implementation, 1996 and 1997. The micro student files included information on achievement measures and student characteristics (gender, parental schooling, family size, immigration status). A student identification number allowed determination of the dropout rate from grades 9 to 12. The school data file included information on schools (size, number of teachers, boarding school facilities) and teachers' characteristics (age, experience, education) in the last year of both

programs (1997). School ID numbers were used to match and merge school files with student files. Whereas the teachers incentives program included 62 schools (37 non-religious, 18 religious and 7 Arab schools), the school resources program included 22 schools (13 non-religious schools, 4 religious schools and 5 Arab schools). There is about 320 high school in Israel, of which 170 are comprehensive high schools.

III. A Graphical Presentation of the Identification Strategy

The ideal strategy for evaluating the effect of the teachers' performance incentives intervention (or any similar intervention) would involve the random assignment of pupils to either treatment or control groups, with those in the treatment group being taught by teachers who stand to benefit from performance bonuses. Random assignment insures that pupils in the control group are indeed comparable to pupils in the treatment group, so that any post-experiment difference between pupils in the two groups could be confidently attributed to the intervention. In the absence of random assignment, statistical methods must be used to control for differences between pupils in treated and non-treated schools.

Table 1, panel A, columns 1 and 4, present the mean of achievement measures and students' and school characteristics of the treated and all the non-treated non-religious secondary comprehensive schools in Israel. Clearly, the two samples differ considerably in all dimensions of outcomes and characteristics. For example, the mean average score in the treated schools (line 3) is 73 and in all other schools it is over 78. The t test for the equality of means is 14.9 (column 5), indicating that they are different at conventional level of significance. Large and significant differences are also evident in student characteristics. These differences suggest that the student body in treated schools comes from a lower socio-economic background (lower parental schooling, larger family size). The comparison of means also indicates that the treated schools are smaller and have younger teachers who have less experience and education. It could also be that the two groups differ in other, unmeasured dimensions, which render the comparison of the two groups

useless for evaluation.

However, there might be a subgroup of schools that could be an appropriate comparison group for identifying the effect of the intervention. The selection rules into the teachers' performance incentives project suggest a natural way to select such a comparison group. Two rules were used to select participants: First, the school must be the only one of its kind in the community. Four types of schools were considered: secular Hebrew secondary schools, boys' and girls' religious secondary schools, and Arab secondary schools. The distribution of the number of schools in the community of each of these four types of schools was used to select potential schools for treatment, namely, those of which there is only one in the community. The second rule, being a comprehensive secondary school, was then applied to the potential list of schools. Therefore, a school was selected to the program if it was a secondary comprehensive school and the only one of its kind in its community, but not if it was one of two or more schools of its kind in the community.

It is useful to present the rationale of this identification strategy graphically. Figures 1-6 present the 1993 and 1994 pre-program averages of all achievement measures for the distribution of schools by the number of school alternatives in the community (for all secular schools). The horizontal axis measures the number of schools in the community, starting from one school in the community, then two, and then averages for the following groups: 3-4, 5-6, 7-8, 9-10 and 11+ schools in the community.

All the figures show that the mean achievement increases with the number of schools in the community (in some of the achievement measures the trend assumes a monotonic step function). This trend probably reflects the correlation between the size of the locality and variables that affect school quality, such as family and community wealth or parental education, as well as the effect of the level of competition among schools on school quality (assuming that the level of competition increases with the number of schools in the locality). We also observe in all figures that the mean achievements of the one-school-per-community

group are higher than those of the two-schools-per-community group. However, when only secondary comprehensive schools are included (the second admission rule used for selection to the program), the differences between them vanish almost completely. These results are seen in Figures 1-6, where T denotes the treated schools and C denotes the comprehensive schools, which are part of the two-schools-per-community group

Column 2 in Table 1 presents the means of the comprehensive schools that comprise the group of two schools per community, which should be compared to column 1. The striking similarity between the two groups adjacent to the cutoff point suggests that it is appropriate to choose the group of secondary comprehensive two-schools-per-community group as a comparison group for identification. This identification strategy can be viewed as a natural experiment or as an example of treatment being determined by a known discontinuous function of an observed covariate (the number of schools in the community). Campbell (1969) was first to consider the later strategy, denoting it as a regression discontinuity (RD) design. Angrist and Lavy (1999) formulated the RD design in terms of an instrumental variable estimation, an approach also applied in Lavy (1998), and, in a different version, by van der Klaauw (1996). In the current application there is only one point of discontinuity, or only one threshold value: the one between belonging to the group of one or two schools per community. Therefore it is appropriate and sufficient to use only the discontinuity sample, namely those schools who are nearest to the threshold or the point of discontinuity, without having to extrapolate and use the whole distribution of schools. The question that arise in this regard is why not use all the schools in the data for estimation, perhaps select a matched sample based on observables and even control for the number of schools in the locality. The argument against this identification strategy is that the very large differences in abservables between the treated schools and all other schools may suggest also large differences in unobservables. The potential correlation between treatment and unobservables makes a strong case for preferring the natural

experiment for identification and therefore choosing schools in small towns as a comparison group and then perhaps match on observables.

Figure 1 can be used to illustrate this identification strategy in greater detail. Forty-one secular secondary schools are single schools in the community. The mean number of credit units in these schools is not equal to the mean of the two-schools-per-community group. Of the 41 one-school group, all the 37 secondary comprehensive schools were selected to the program. Of the 18 secondary schools in two-schools group, only 8 are comprehensive schools. Comparing the first group's 37 treated schools with the second group's 8 schools reveals striking similarities in mean achievement: the mean of credit units is almost identical (19.3), compared with 21.9 for all non-treated Hebrew secular secondary comprehensive schools in Israel. The similarity between these two 'adjacent' groups of schools is replicated in all the other achievement measures. The t tests of the means (Table 1) indicate that for all the six achievement measures the differences between the means of the two groups are barely significantly different from zero even though the sample is very large. This stands in sharp contrast to the t-test statistics for equality of means between the treated one-school group and all other non-treated schools (column 5).

The middle panel of Table 1 compares the mean demographic characteristics of the students in the two groups of schools: mother's and father's years of schooling, gender and immigration status. There are some small but significant differences, particularly with respect to differences between the treated schools and all other schools in the country. For example, the difference between mother's and father's schooling between the one- and two-school groups is less than half a year; the percentage of immigrant and female students is almost identical. Similarities between the two groups are also evident in school characteristics (school size, number of teachers) and in teachers' characteristics (age, experience, and education).

Using the 1993 instead of the 1994 data leaves this result unchanged: the group of treated 'single' schools is almost identical in achievement measures to the group of untreated 'two' schools (Figures 1-6)

in the community. Likewise with regard to student and school characteristics. The stability of the pattern in both pre-treatment years may be an indication that these similarities are permanent, enhancing the credibility of the identification model. These similarities can also be viewed as a reasonable justification for a differences in differences estimation method to which I turn next.

Differences-in-differences estimates

Figures 7a-7f provide the basis for simple differences-in-differences (DID) estimates of the effect of the teachers' incentive program (non-religious schools) on the various measures of achievement. For example, the 1994 averages of the treatment and control groups can be compared to the post-program averages of the two groups. Using the 1994 (pre-program — solid lines) and the 1997 (post-program — dotted lines) averages, these figures contrast the pre and the post means of both groups for the six achievement measures. The differences between the solid and the dotted lines represent the absolute change in achievement. The differences in these absolute changes for both groups are the DID point estimates of the average treatment effect of the program. Contrasting the 1996 with the 1994 means yields DID estimates that are practically zero for most achievement measures and therefore they are not presented in the figures.

The 1997 DID estimates are positive for all achievement measures except for the proportion of pupils who earned matriculation certificates.[11] For example, the change in the mean number of credit units is 1 for the treated and 0.6 for the control schools. The difference between these two changes is 0.4 units. The mean change in the average matriculation score is about 1 point for treated schools and 0.05 points for control schools, the DID estimate is 0.95 points. The DID estimate for units of science credits is 0.15 and for the proportion of pupils taking matriculation exams — 0.02. However, all these estimates are not precise. Furthermore, since the means compared are for different cohorts of students, the change estimated may

[11] The proportion of pupils who earned matriculation certificatesis unconditional on their having taken the exam.

merely reflect student compositional change. The next section presents the methodology that allows us to study the DID estimates net of potential compositional change and of the effect of unobserved school heterogeneity, using a panel of micro student-level data.

Treated Hebrew religious schools

The sample of treated Hebrew religious schools belongs to the one-school-per-community group, but the data include many other schools that belong to the same category but were not included in the incentives program (14 Hebrew religious schools). While Ministry of Education officials argue that all the schools that satisfied this condition in 1995 were included in the program, it could very well be that the Ministry failed to classify these other schools properly as belonging to the one-school-per-community category. This bureaucratic error raises the possibility that the group of 'misclassified' schools is a natural comparison group for the identification of the treatment effect of the teachers' performance incentives in Hebrew religious schools.

Figures 8-14 present the 1993 and 1994 pre-program averages of all achievement measures for the distribution of schools by number of school alternatives in the community in the Hebrew religious school system. The scale starts from one school in the community, then two schools, and then averages for the following groups: 3-4, 5-6, 7-8, 9 and more schools in the community. Again we observe that mean achievement increases with the number of schools in the community, though the shape of the step function is not as monotonic as in Figures 1-6, probably due to the small number of schools in each cell. It is also seen in all figures that mean achievements in treated schools are lower than the means of the non-treated one-school-per-community group (and also of the two-schools-per-community).

Panel B of Table 1 presents the means of the religious schools. Column 1 presents the means for the treated schools. Column 2 presents the means of the comprehensive schools forming the group of untreated

one-school-per-community, which should be compared to column 1. The means in column 2 are much more similar to those of column 1 than the means in column 4, suggesting that the group of untreated secondary comprehensive one- school-per-community is a more appropriate comparison group for identification.

IV. Measurement Framework

Figures (7a-7f) suggest a clear and positive relationship between the exposure of teachers to performance incentives and post-program students' achievement in the second year of the program. This evidence is based on a comparison of post-treatment outcomes in participating schools and the natural experiment (RD) comparison schools. Based on the very similar means of the pre-treatment achievement measures and characteristics, the maintained assumption was that the two groups of schools are identical in all pre-treatment measures. However, since pre- and post-treatment school-level panel data are available, we can extend the simple model to include school fixed effects that will account for any remaining permanent differences, observed and unobserved, between the two groups.

This extension is simply an application of a model used in numerous evaluation studies (e.g., Ashenfelter, 1978; Ashenfelter and Card, 1985; Card and Sullivan, 1988), which is based on the assumption that, in the absence of actual treatment, any differences between treatment and comparison schools are fixed over time. In this case, repeated observations of schools, though their constituent students change every year, can be used to account for any remaining differences between the two groups.

Let D_{ist} be a dummy variable that indicates treatment status and let $Y_{ist}(0)$ denote the potential achievement score of any pupil in school s at time t if his or her school was not exposed to the treatment. The fixed-effects model implies that in the absence of the intervention the potential achievement score of pupil i in school s (treated and comparison schools) at time t (1993 through 1997) can be written as

$$Y_{ist}(0) = X_{ist}'\beta + \phi_s + \delta_t + \varepsilon_{ist}\ , \qquad (1)$$

where ε_{ist} is a time-varying error term and ϕ_s is a school-specific intercept. ε_{ist} can be correlated with ϕ_s but it is assumed to be independent of the treatment status D_{ist}. The term δ_t is a period effect common to all pupils and schools. Thus, while pupils in the treated schools may still have higher or lower pre-treatment achievement measures than pupils in the control schools, this difference is assumed to stem from differences in family background and school characteristics (X_{ist}) or from unmeasured school characteristics (ϕ_s) that can be viewed as permanent and that have time-invariant effects on test scores.

Starting with a simple model that also has a constant treatment effect, treated pupils are assumed to have achievements measures equal to $Y_{ist}(0)$ *plus* the treatment effect:

$$Y_{ist}(1) = Y_{ist}(0) + \alpha\ . \qquad (2)$$

Equations (1) and (2) can therefore be used to write the observed achievement score of pupil i in school s at time t as

$$Y_{ist} = Y_{ist}(0)[1 - D_{ist}] + Y_{ist}(1)D_{ist} = X'_{ist}\beta + \phi_s + \delta_t + \alpha D_{ist} + \varepsilon_{ist}\ , \qquad (3)$$

where ε_{ist} is the error term in (1), assumed to be uncorrelated with D_{ist}. Note that D_{ist} in (3) is an interaction term equal to the product of a dummy indicating observations after 1994 and a dummy indicating treated schools. In this model, simple post-treatment comparisons of test scores by treatment status do not estimate the causal effect of treatment because the time-invariant characteristics of schools in the treatment and control groups differ. However, we can use the panel data to estimate ϕ_s and obtain a consistent estimate of the program effect. This estimate is basically the difference in differences estimators computed in a regression of stacked students (micro data) for schools and years, when a vector of students'

characteristics is added to the regression.[12]

The evidence in Table 1 indicates that there are still some differences between pupils in the two groups: pupils in the treatment schools have lower pre-treatment achievement measures, on average, than pupils in the control schools. The above model controls for these pre-treatment differences by assuming that they are fixed at the school level and subtracts them from the post-treatment difference in scores. This assumption seems appropriate given the stability of the differences as seen in the two years prior to treatment. However, instead of relying solely on this model, a matching strategy can be used to non-parametrically control for any remaining average pre-treatment achievement differences (see, e.g., Heckman, Ichimura, and Todd, 1997). The matching can be based on propensity score analysis (see, e.g., Dehejia and Wahba, 1998) or on school characteristics (see, e.g., Angrist and Lavy, 1998, and Angrist, 1998). In this paper, since the number of schools in each group is small, the matching is accomplished by limiting the sample to treatment and control schools where school-level *average* test scores are similar in the two groups. This is useful if there is something special about the treatment schools ("random school effects") that cannot be captured by the RD design and the estimated school fixed effects.

The school fixed effect models assume that the differences between treatment and control schools are permanent. If some of the differences are temporary, owing to transitory shocks, the fixed effect estimates will be biased. Two comments should be made in this regard. First, we have seen that the similarities between treatment and control schools hold for each of the two years prior to treatment, 1993 and 1994. Second, the effect of the two programs is evaluated over two or three years of implementation, which reduces the possibility of a 'regression to the mean' effect.

[12] This statistical model is identical to the one used by Card (1990) in his study on the effect of the Mariel boatlift on employment in Miami.

Identifying the effect of school resources

Evaluating the effect of the school resources program is based on using applicants who were not admitted to the project as a comparison group. Such a comparison controls for differences between participating and non-participating schools that originate in the decision to apply to the program, but does not control for most of the criteria used by the authorities when choosing which of the applicant schools to accept. As shown in the following section, the applicants are much more similar to the treated schools than to all the other schools, but there are still some remaining differences in measures of achievements and characteristics. Again, using the panel data on schools to estimate equation (3) will account for these and other (unobserved) differences, as long as they are permanent. Assuming that the treatment status among applicants can be ignored, conditional on the set of observed students covariates and the school fixed effects, this comparison will yield a consistent estimate of the program's effect on the treated students.

As in the case of the teachers' incentives intervention, a non-parametric school matching strategy can be used to achieve even better pre-treatment observables equivalency between the two groups. The approach is similar to Angrist (1998) who used a matched sample from applicants to the military who did not serve in the army as a comparison group vis-á-vis applicants who did serve in he military.[13] As in the previous intervention, each of the treated schools is matched to an untreated applicant with the same 1994 proportion of pupils who received matriculation certificates. This matching controls for pre-program temporary shocks leading to transitory differences between schools that influenced participation in the project. Using two years (1993 and 1994) instead of one as a base for the estimation of the school fixed effects should help to eliminate the temporary from the permanent differences between schools.

[13] Angrist (1998) actually matches individual applicants to the military to veterans, based on individual charactersitics; here I match applicant schools to treated schools like in Angrist and Lavy (1998).

V. Empirical Results

a. *The teachers' incentives intervention*

Tables 2 and 3 present the estimated treatment effect for secular and religious Hebrew schools, respectively, in 1996 and 1997. For each year three different panel data sets were used to allow for school fixed effect estimation. The first line in each panel presents the results from a panel that includes the two pre-treatment years and one treatment year, 1996 or 1997. The second and third lines in each panel present the results when only one pre-treatment year was included, 1993 or 1994.

The first rows in panels A and B in Table 2 suggest that intervention in the form of teachers' incentives had a positive and significant effect on student outcomes in the second year of the program, but only a minor effect in the first year. The evidence is relatively robust with respect to which year is used as the pre-treatment period. The estimated effects in the second year are positive and significantly different from zero on all achievement measures except for the proportion of pupils who earned matriculation certificates. Focusing on the model that uses both pre-treatment years as a base (the first row in panels A and B in the table), the program in 1997 is seen to have increased average credits by 0.70 units (SE = 0.34), science credits by 0.41 units (SE = 0.15), average score by 1.75 points (SE = 0.85), and the proportion of students sitting for the matriculation exams by 2.1 percent (SE = 0.013).[14] The estimated effect on the proportion of students earning a matriculation certificate is essentially zero.

The same model was estimated with a sample that included only matched schools. I adopt a heuristic matching strategy since there are several measures of achievement and only a few schools to match. Each of the eight control schools was matched to a treated school with the same 1994 average matriculation certificate-passing rate. As a result, only 8 of the 37 treated schools (and all the 8 control schools) now

[14] The standard errors for the regressions estimates reported in all the tables in the paper are corrected for school-level clusetring using equation (1) in Moulton (1986).

appear in the analysis. Naturally, the two samples become even more similar in achievement and characteristics. The 1997 matching results (the even columns in Table 2) are qualitatively very similar to those obtained with the full regression discontinuity sample. The effect of the program on science credits units and average scores is somewhat stronger in the matched sample. Using the matched sample improves the results regarding the effect of the program on the proportion of pupils entitled to a matriculation certificate: this effect is estimated as .028 with a standard error of .022.

Using 1994 or 1993 as a base in the panel sample instead of both years together leads to some variations in the results, suggesting that there is some sensitivity to the choice of the base year, but the differences in the estimated effects are not dramatic. Since there are random or transitory school effects that may lead to an effect of convergence to the mean in the following year, it is better to use more than one year as a base period.

The results for religious schools show that the program had a significant positive effect on all measures and in both years of the program. Minor differences in the size of the estimated effects are observed between the full and the matched samples, but overall the two sets of samples and the two years yield consistent estimates. The relatively minor sensitivity to which year is used as base year is also evident in the religious schools results. The effect of the program on the number of science credit units is much smaller and less precisely estimated in the religious schools sample compared to the secular schools sample. This is not surprising, given that religious schools, on average, place less emphasis on science subjects.

The effect of the program in the religious schools is larger then in the secular schools. Using for comparison the results obtained from the RD sample and 1993 and 1994 as base year, the absolute effects can be expressed as a ratio of the means in the sample. The respective ratios for the religious and secular samples are 6.3% versus 3% for credit units; 4.1% versus 2.4% for the average score; 12.4% versus 4.2% for the proportion of pupils taking the matriculation exams.

Which students benefited from the program?

The results presented above suggest that the mean number of credit units and the average score increased but that the proportion of students of qualified for the matriculation certificate did not change on average. These two pieces of evidence can be consistent with each other under two scenarios. The first is that the treatment effect is mainly on good students who would have qualified for matriculation regardless of the program. The second scenario is that the main effect of the program is on weak students who are still far, in terms of number of units and average score, from the threshold of gaining matriculation. A simple way to trace the composition of students who were affected by the program is to add to the model interaction terms of the treatment variable with the demographic variables. In all the models presented above, these variables (father and mother schooling and family size) had very large and significant effects on all the outcome measures. It is clear from these results that the low end of the distribution of students, in terms of outcomes, is characterized by low level of parental schooling and large family size.

Table 4 presents results for 1997, for secular and religious schools, from models that included treatment interactions with father and mother schooling and with the number of children in the family. The results clearly suggest that the treatment effect in all the five outcome measures decline with all the three demographic variables, especially with mother's schooling. The coefficient on the interaction with mother's schooling is significantly different from zero in all five cases, in the secular and religious samples. Two important conclusions can be based on Table 4. First, the teachers incentives program affected mainly the weak students: at the sample mean of mother's schooling the net program effect is almost zero. Second, the intervention led to an increase in the rate of students who achieved the matriculation certificate among students from poor socio-economic background even though on average it did not have any effect on this rate.

The estimated effects on dropout rates

The incentives program aimed at reducing dropout rates as well. The program focused relatively more on reducing the dropout rate in the transition from 9th to 10th grade. This transition involves the graduation from middle school and enrollment in an upper secondary school. However, I will also report below results from estimating the program effect on dropout rate from 10th or 11th grade and from 11th to 12 grades.

To evaluate whether the intervention had any significant effect on high school dropout rate, I estimated probit models with school fixed effects, similar to the models discussed above. The dependent variable was an indicator of whether the pupil dropped or continued schooling in any school in the country. I will first report estimates of the program effect on the school continuation rate from 9th to 10th grade. Since data on 9th grade students is available only from 1994 on, all the models reported below are estimated using 1994 as the pre-treatment base year. The 1996 secular school sample included 7,795 pupils who completed 9th grade. The respective sample size in 1997 is 5511 pupils. The mean dropout rate in 1996 is 6.0 and in 1997 it is 5.5 percent. The religious schools 9th sample included 3,157 pupils in 1996 and 2,290 pupils in 1997. The mean dropout rate in 1996 is 3.7 and in 1997 it is 3.2 percent. Table 5, columns 1-4, presents the estimated effects of the incentive intervention on dropout rates at 9th grade. The effects are negative and relatively precisely estimated, especially in the secular schools sample. Like the results presented above for the effect of the teachers' incentive intervention on scholastic achievements, the estimated effect on dropout rate is improved in the second year of the program (1997) compared to its effect in the first year (1996). For example, the estimated program effect in column (1) of Table 5 is -.253 (s.e = .123) in 1996 and -.600 (s.e = .159) in 1997. The estimated effects are larger and with smaller standard errors when the matched samples are used. In the religious schools the estimated effects are negative but they are precisely measured only when the matched sample is used. These estimated effects are even larger then their

counterparts obtained from the secular matched sample (column 2 versus column 4 in Table 5).

Columns 5-8 present the estimated program effect on the continuation rate from 10th and 11th grade. The probit regressions were estimated using a pooled sample of pupils from these two grades. The 1996 secular school sample included 30,784 pupils who completed 10th or 11th grade. The respective sample size in 1997 is 31,296 pupils. The mean dropout rate in 1996 is 6.2 and in 1997 it is 4.8. The religious schools 10^{th} to 12th grade sample included 10,328 pupils in 1996 and 10,634 pupils in 1997. The mean dropout rates for these two years are 4.0% and 2.7%, respectively. A dummy variable is used in the regressions to distinguish the 10th graders from the 11th graders in the sample.

The estimated effects the incentive intervention on dropout rates are negative but very imprecisely estimated (Table 5, columns 5-8), especially in the religious schools sample. The estimated effect is not improved in the second year of the program (1997).[15] These evidence suggest that the main effect of the program on drop out rates is on continuation rates from the last year of middle school (9th grade) to the first year of high school (10th).

Characterization of 'winning' schools

Although the teachers' incentive program was introduced to teachers and principals in the treated school only in middle of the 1994/95 school year, bonuses were given for this year as well, mainly in order to publicize the program and enhance its credibility. Eighteen schools received bonuses for this year (11 of the 37 participating secular schools and 8 of the 19 participating religious schools). In the full two years of the project, 1995/96 and 1996/97, 13 and 17 secular schools received bonuses, respectively. The number of award winning religious schools remained 8 in both years. It should be noted that the religious and the

[15]Data for high school students is also availble for 1993. Using this data to add 1993 as pre-treatment base year in the estimated models did not change the results in any signinficant way.

non-religious schools were pooled together for determining the ranking of schools that served to determine the best performers who won financial bonuses.

Seventeen of the 56 participating Hebrew schools did not win an award in any of the three years. Nineteen schools won once, 9 won twice and 8 won in all three years. Lazear and Rosen (1981) analyzed compensation schemes which pay according to an individual's ordinal rank in a tournament rather than his output level. In such a tournament model with symmetric information, risk neutrality and equal abilities of individuals (schools), all individuals should make the same effort and the outcome is a random event, namely the probability of winning is equal to the proportion of schools who receive prizes. In our case this proportion is a third and therefore the probability of winning twice in a row is 0.11. Focusing on the two years in which the intervention could have had a causal effect (1996 and 1997), namely in the years in which schools and teachers were fully aware in advance of the competition, 14 of the 53 Jewish school who participated won twice. This proportion, 0.28, is much higher then 0.11 predicted from the above model. These results suggest that there is something systematic and not random that generates the results.[16] More information pointing to this same direction can be derived by comparing the proportion of schools who won both in 1995 (no intervention but prizes distributed) and 1996 (first year of treatment) to the proportion of schools winning in 1996 and 1997. We expect the proportion of winners in both 1995 and 1996 to be much smaller than the proportion of winners in both 1996 and 1997. Focusing only on the secular schools who had a relatively large sample of participants, 4 of the 36 schools (11 percent) won in 1995 and 1996 while 9 of the 36 schools (25 percent) won both in 1996 and 1997.

Comparison of award-winning schools in 1996 or in 1997 to the non-winning schools in each of these

[16] Some schools may decide in the beginning of the tournament that their chances of winning is nill and therefore they make no real effort. In such a case the "actual number of participants" is actually smaller and therefore the number of winners is from a smaller number of players. This kind of behavior can be an alternative explanation of the high proportion of schools that won twice in a row.

years suggests that they do not differ in any of their pre-treatment, 1993 or 1994, characteristics. The students' scholastic achievements in the two groups (winners and non-winners) are about the same, in 1993 and 1994. Similarly for students' characteristics (parental education, family size and percent male and immigrants students) and school characteristics (school size, number of teachers, boarding school). No differences in the 1997 teachers' characteristics emerge from this comparison as well.

b. The school resources intervention

The school resources program included much fewer schools than the incentives program. Only four religious schools were included in the former program. One of these schools had to be excluded from the sample because of missing data for some of the students' variables leaving too few treated religious schools for a meaningful evaluation. Therefore the resources intervention is evaluated here only for non-religious schools. Table 6 presents the means of measures of achievement and characteristics of students and schools included in the treated non-religious schools (column 1). Also included in the table is the same information for all other applicants to the program (column 2) and a matched sample from this group (column 4). Overall, students from non-treated applicant schools show higher achievements than students in treated schools. These differences, though significantly different from zero in most cases, are not very large in absolute terms. However, in order to eliminate these differences as well, a matched sample was chosen for comparison: to each of the treated schools I matched one of the not-treated applicants who had the same school average 1994 proportion of pupils who received the matriculation certificate. As a result, the gap between the means presented in columns 1 and 4 are smaller than those presented in columns 1 and 2, although some differences remain.

Table 7 presents the estimated coefficients of the resources intervention in secular schools. The odd columns report results from the full set of applicants and the even columns — those from the matched

sample. The first three rows in parts A, B and C of the table present results of variation in the years used as the pre-treatment base period. The estimated effects of the program in its first year are minor, being positive and significantly different from zero only in the regressions of credit units and average score. The estimated effects for the second and third year of implementation are more impressive, being positive and significantly different from zero for achievement measures, except for the proportion of students who earned matriculation certificates. The sensitivity of the results to the years chosen as a base period is similar to the one revealed in the incentives' intervention (Table 2).

The applicant schools were screened and a group of 32 schools out of the 70 applicants were recommended for a second review and final selection. These 32 schools presented their program to a committee that ranked them in terms of quality[17]. This list of recommended schools could be used as a control group instead of all the applicants. Table 8 presents the results obtained when the non-treated recommended schools was used as a control group instead of all the applicants. Focusing again on the 1997 results for comparison suggests that the use of the recommended schools as a comparison group leads to higher treatment effects: 2.266 versus 1.066 for credit units, 5.338 versus 2.897 for average score, .088 versus .084 for the proportion of students taking the matriculation exams and .027 versus -.002 for the proportion of pupils who earned the matriculation certificate.

The question of which students benefited from the intervention is also relevant in the resources project. Table 9 presents estimation models that included interaction of the treatment variable with the student demographic variables. In all cases the interaction terms of treatment with parental schooling are negative, but in most cases they are not significantly different from zero. However, the interaction with family size is negative also, and in most cases it is large and significantly different from zero. The negative sign of this

[17] Actually, four commitees reviewed 8 schools each and ranked them separately. The ranking of one of these groups is lost, rendering the ranking data usesless for analysis.

variable indicate that the treatment effect is larger for students who come form small families, a status highly correlated with low income and with low educational achievement measures. This result suggest that the resources program, unlike the teachers' incentive program, had most of its effect on the better students, those who are above average in their performance measures.

The effect of the resources program on dropout out rates is presented in Table10, columns 5-8. The estimated effects on dropout rates from 10^{th} or 11^{th} grade are all negative but they are very imprecisely estimated. Only in the 1997 sample of applicant schools that were recommended to the program the treatment effect estimates are significantly different from zero at the 5% level of significance. The 1997 estimate is -0.199 (SE=0.087).

VI. The Relative Cost-Effectiveness of the Two Interventions

A comparison between the 37 secular schools included in the teachers' incentive project and the 13 secular schools included in the resources project reveals that both groups are very similar in pre-treatment achievement measures and characteristics. Comparing column 1 in Table 1 with column 1 in Table 5 points to striking similarities. For example, the mean number of credit units in the first group is 19.2 and in the second it is 19.3. The respective figures for average scores are 74.7 and 73, and for matriculation certificates — 0.48 and 0.47. Even the mean dropout rates are identical: 6 percent. Student characteristics are also very similar, especially with respect to parental schooling. Similar results are obtained from a comparison of the school and teachers' characteristics of the two groups. All these similarities, which are purely coincidental, make the comparison of the two intervention alternatives more meaningful.[18]

[18] The comparison of the effects of the the two interventions requires that the two comparison groups also be similar in pre-treatment characteristics. Comparing columns 2 in Tables 1 and 5 reveals differences between the two groups but they are not very large. But comparing the matched comparison group in Table 5, column 4, to the comparison group presented in Table 1, column 2, suggests that the two groups are more similar.

The relevant comparison between these two policy options is the one based on cost equivalence. The total cost of the teachers' incentives program in 1996 was higher: $1.44 million annually versus $1.13 million annually for the school resources program. However, the incentives program affected almost triple the number of schools than did the resources program (62 versus 22 schools). Therefore, the cost *per school* of the resources program ($51,600) was more than double the average cost per school in the incentives program ($23,300). To determine which of the two programs was more efficient, the gap in the cost per school should be weighed against the difference in the effect of the two programs on output.

A comparison of the results in Tables 2 and 5 suggests that in most of the achievement measures the estimated effect of the resources program is greater than that of the incentives program, but this advantage is not sufficient to offset the cost advantage of the incentives program. Focusing for simplicity of exposition on the results for 1997, which are based on using 1993 and 1994 as base years, the resources program had a greater effect in three of the four measures of achievement. These results are not sensitive to the use of the full or matched samples in the analysis. Using the full sample results, for illustration, the effect of the resources program on credit units is 50 percent higher (1.16 versus 0.76). The gap is 40 percent on average scores (2.9 versus 2.0) and 60 percent in the proportion of pupils sitting for matriculation exams (0.084 versus 0.031). All these differences are significantly different from zero at conventional levels of statistical significance. However, the incentives program had a significant effect on the number of science credit units, while no effect on this outcome is estimated for the resources intervention. Both programs had no effect on the proportion of pupils who earned matriculation certificates.

Our results indicate that the resources program had on average a 50 percent higher effect on outcomes than did the incentives program, but cost more than twice as much. Therefore, per marginal dollar spent, the teachers' incentive intervention seems to be much more cost-effective. The analysis and result do not change if the cost comparison is done on a per-student basis since the mean number of students in schools

in the incentives' project (1,076) is similar to the size of schools in the resources project (1129). The above cost-benefit analysis is valid under the assumption that the technology linking the interventions to output is not the same for the two programs. If the same function, relating allocated money to outcome, could describe the two programs then the above analysis will be valid only under the assumption of constant return to scale. When there are decreasing returns to money allocated to school, then the observed pattern is possible even in the case in which money allocated to incentives and resources is a perfect substitute. In this case a more expensive program will have a lower outcome per dollar spent, regardless of whether it is an incentives or resources program. However, the assumption that the technology relating money to outcome is different for the two programs seems more realistic in this case.

Several other aspects of the two programs should also be relevant for interpreting the results of the cost-equivalency comparison. (a) The effects of the resources program could be felt already in its second year, whereas the incentives' intervention had a very limited effect in its first year, especially in the non-religious sample. But since the incentives program started a year later than the resources program, comparing its results to those of the resources program in its third year is hardly fair. (b) The boost received by schools through the resources program amounted to a significant increase in total school teaching resources (on average more than two teachers per school), or about 1.3 percent of the average number of teaching staff in the treated schools. On the other hand, the salary bonuses awarded to teachers were very small relative to their annual income: the highest bonus per teacher was $715 and the lowest $200. Since average gross annual income of teachers is $30,000, the maximum bonus is about 2.5% of annual income while the lowest bonus is less than 1% of annual income.

VII. Summary and Conclusions

The idea of developing stronger performance incentives directly focused on students' achievements has vast

appeal, and is the subject of frequent discussion. But incentives have seldom been tried in practice, and experiments with them are even more limited. Studies and analyses of the few available examples lack many of the details crucial to judging their more general applicability. In this paper I present an evaluation of the effect on students' achievements of a teachers' performance-based incentives program implemented in Israel in a large number of schools. The program had the main elements of a rank order tournament where only the top third performers were awarded monetary bonuses. The rules of selection of schools into the program present an opportunity to base the identification strategy on a natural experiment. The results suggest that the teachers' performance incentives led to significant gains in four out of five achievement measures of high school graduates, including average test scores and the number of science and other credit units, and on the dropout rate from 9th to 10th grade. The program did not lead to an increase in the proportion of students who qualified for matriculation certificates and to a decline in dropout rates. These results were obtained for both parts of the Hebrew school system, religious and non-religious schools. Since the program was implemented at schools in relatively small communities we should be cautious in extrapolating the results to other environments.

An alternative to the program based on incentives — more school resources, in the form of teaching time and focusing on potential dropouts and weak students — also had a significant effect on students' achievements. This result should come as no surprise, since this program, too, encouraged the creativity and effort needed to develop and implement effective interventions. This was done by giving schools complete control over the additional resources and total freedom to shape its elements. This is a feature common to both programs and perhaps it is effective. However, a comparison based on cost equivalency indicates that the teachers' performance-based incentives program is more efficient.

The teachers' incentives program studied in this paper is only one of many possible variations. Little is known about what forms of incentive systems are best in general or specific circumstances, nor about

precisely what results might be expected from making wider use of any specific system. However, the importance of the evaluation presented here lies in the fact that the power of incentives observed elsewhere in the economy is also evident in schools, even in the case of relatively low performance bonuses.[19] Finding the best ways to structure incentives in schools, for teachers and students alike, should be based on more experimentation and evaluation.

[19] The list of schools that win the tournament is announced every year bu the Minsiter of Education in a press conference. The large media exposure that these schools receive folloeing the anouncment may compensate for the low level of the monetary awards.

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): 1-44.

Angrist, J. and Lavy, V. (1998). "Does Teacher Training Affect Pupil Learning? Evidence From Matched Comparison in Jerusalem Public Schools." NBER Working Paper 6781.

Ashenfelter, O. A. (1978). "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics* 60 (1): 47-57.

Ashenfelter, O. A. and Card, D. (1985). "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs on Earnings." *The Review of Economics and Statistics* 67 (4): 648-660.

Campbell, D. T., "Reforms as Experiments," *American Psychologist* 24 (1969), 409-429.

Card, David E. (1990). "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review*, 43:245-57.

Card, D. and Sullivan, D. (1988). "Estimating the Effect of Subsidized Training on Movements In and Out of Employment." *Econometrica* 56 (3): 497-530.

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.

Dehejia, Rajeev H. and S. Wahba. (1998). "Causal Effects in Non-Experimental Studies: Re-evaluating the Evaluation of Training Programs." NBER Working Paper # .

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.

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

Heckman, J. J., (1998) "Understanding the Sources formation In A Modern Economy: models, Evidence

and Implications for Policy." Draft of a lecture at Tel-Aviv University, December 16, 1998.

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.

Hanushek E. A. (*et al.)* (1994). *Making Schools Work: Improving Performance and Controlling Cost*. Washington D.C.: The Brookings Institution.

Hanushek E. A. and D. W. Jorgenson (eds.) (1996). *Improving America's Schools*, National Academy Press, Washington, D.C., USA.

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.

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.

Kandel, E. and E. Lazear (1992). "Peer Pressure and Partnership." *Journal of Political Economy* 100 (4):801-17.

Lavy, V. (1998). "Using Quasi-Experimental Designs to Evaluate the Effect of School Hours and Class Size on Student Achievement." Mimeo., The Hebrew University of Jerusalem, Department of Economics.

Lazear, E. (1996). "Performance Pay and Productivity." NBER Working Paper 5672.

Lazear, E. and S. Rosen. (1981). "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy* 89: 841-64.

Milgrom, P. and J. Roberts (1992). *Economics, Organization and Management* , Prentice Hall, New Jersey.

Moulton, Brent, "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics* 32 (1986), pp. 385-97.

Van der Klaauw, W. (1996). "A Regression-Discontinuity Evaluation of the Effect of Financial Aid Offers on Enrollment," unpublished manuscript, New York University.