Teachers' Pay for Performance in the Long-Run: The Dynamic Pattern of Treatment Effects on Students' Educational and Labor Market Outcomes in Adulthood*

Victor Lavy

University of Warwick, Hebrew University of Jerusalem and NBER

March 2019

Abstract

This paper examines the dynamic effects of a teachers' pay for performance experiment on long-term outcomes at adulthood. The program led to a gradual increase in university education of the treated high school students, reaching an increase of 0.25 years of schooling by age 28-30. The effects on employment and earnings were initially negative, coinciding with a higher rate of enrollment in university, but became positive and significant with time. These gains are largely mediated by the positive effect of the program on several high school outcomes, including quantitative and qualitative gains in the high-stakes matriculation exams.

*v.lavy@warwick.ac.uk. Excellent research assistance was provided by Boaz Abramson, Elior Cohen, Michal Hodor and Genia Rachkovski. I thank Jo Altonji, Josh Angrist, Peter Dolton, Zvi Eckstein, Nathaniel Hendren, Ed Lazear, Yona Rubinstein, Uta Schonberg, three referees of this journal and participants at seminars at Hebrew University, University of California Santa Barbara, IDC public Economics Annual Symposium, Warwick-Venice 2016 Labor Economics Conference, Barcelona 2016 Summer Forum and CESifo Economics of Education 2016 Conference for useful comments and suggestions. I thank Israel's National Insurance Institute (NII) for allowing restricted access to post-secondary education and economic and social outcomes data at adulthood in the NII protected research lab. I acknowledge financial support from the European Research Council through ERC Advance Grant 323439, the Israeli Science Foundation and the Maurice Falk Institute in Jerusalem.

1. Introduction

Skeptics allege that teachers' pay for performance (PFP) schemes only improve student test scores by pushing teachers to teach to the test, or by encouraging teachers and schools to cheat. They claim there is no real increase in human capital because teachers do not respond to pay incentives by promoting broad human capital acquisition. For example, teachers can narrow the focus of their teaching and only include material in the exam, teach students how to take tests, demonstrate exam solving strategies, and instill skills and actions that raise scores on the formulas used to reward teachers.[1] Concerns about narrowly targeted gains are heightened if those gains are focused on areas where labor market rewards are due to signaling rather than human capital acquisition.

To address these claims, I examine the effect of teachers' PFP on long term human capital outcomes, in particular, attainment and quality of higher education, and labor market outcomes at adulthood, such as employment and earnings. I use a teachers' PFP experiment which I conducted almost two decades ago in Israel. In Lavy (2009) I analyzed the short-term effects of this experiment on students' cognitive high school outcomes. I now use this earlier research to evaluate whether an intervention that offered teachers bonuses for student test achievements has had a lasting impact on adult well-being. This paper provides the first evidence of links between teachers' PFP during high school and students' schooling, labor market outcomes and marriage and fertility in their late 20s and early 30s. Some of these outcomes, for example, tertiary schooling, can also be viewed as potential mechanisms for the effect of the intervention on employment and earnings.

I observe these students' outcomes every year, from high school graduation until age 30 (in 2012). Thus, I can estimate the treatment effects for every year in the period, and trace the dynamic evolution of the program effect. Since a high proportion of the sample was in military service for two (female) or three (male) years after high school,[2] the estimates for these years (2000-2004) are not very informative, because they are based on a small and selective sample of those students not doing military service. However, during 2005-2008 the treated group showed a higher enrollment rate in university education, and a corresponding lower employment rate and lower earnings. By the end of this period, these negative effects were eliminated and the earnings effect turned positive, increasing in size and becoming significantly different from zero 9-11 years after high school graduation.

Just over a decade after the end of the intervention, treated students are 5 percentage points more likely to be enrolled in university and to complete an additional 0.25 years of university education, a 30 per cent increase relative to the control group mean. The most likely explanation for these gains is the improvements in high school matriculation (in Hebrew termed *bagrut*) outcomes facilitated by the

[1] See Jacob and Levitt (2003), Glewwe, Ilias and Kremer (2010), and Neal (2011) for a discussion of this issue.
[2] Israelis begin a period of compulsory military service after high-school graduation. Boys serve for three years and girls for two. Ultra-orthodox Jews are exempt from military service as long as they are enrolled in seminary (Yeshiva); orthodox Jewish girls are exempt upon request; Arabs are exempt, though some volunteer.

teachers' PFP intervention. The higher passing rate and average score in the math and English matriculation exams (Lavy 2009) are also expressed in improvements in average matriculation outcomes, such as matriculation diploma certification (up by 3.6 percentage points) and the overall composite matriculation score (up by 2.8 points). These two outcomes determine admission to university and to selective degrees, such as medicine, engineering and computer science. Other dimensions of the high school matriculation study program that signal quality of schooling also improved, in particular, the number of science credit units, which increased by 25 percent, and the number of subjects studied at the most advanced level, which increased by 5 percent. These high school outcomes are also highly correlated with labor market outcomes at adulthood. These improvements, along with the increase in university education, led to a 1.3 percentage point gain in employment rate, a quarter of an additional month of work per year and to 8-9 per cent increase in earnings at age 28-30. The estimates suggest that the program did not have an effect on average marriage and fertility rates.

Even though this is the first study to provide evidence of the effect of teachers' PFP on student earnings at adulthood, it is still useful to compare our results to the impact of other schooling interventions on earnings at adulthood. Chetty et al. (2011) have shown that having a kindergarten teacher with more than ten years of experience increased students' average annual earnings at ages 25 to 27 by 6.9 per cent. Johnson et al. (2016) show that for children from low-income families, increasing per-pupil spending by 10 per cent in all 12 school-age years increased adult hourly wages by 13 per cent. Clearly, our estimated effects on earnings are not unusually high when compared with these estimates.

These average gains mask some heterogeneity by family income and gender. For example, children from families with above median income experienced a higher increase in schooling but no effect on employment, while children from families with below median income have fewer gains in schooling and a large positive effect on employment and larger effect on earnings. Children from families with below median income experienced a decline in marriage rate and a more modest decline in fertility.

The results of this analysis have meaningful external validity and are easily transferable and applicable to education in other developed countries. The high school system and the high-stakes exit exams in Israel are very similar to those in other countries. Importantly, variants of the teachers' PFP intervention studied here have been implemented in recent years in developed and developing countries. This study contributes to the empirical evidence about the returns to education interventions, creating a useful guide for policymakers. Another important advantage of the evidence presented in this paper is that teachers' PFP is an intervention that can be directly implemented by public policy, whereas evidence based on parameters such as school or teacher quality is not so easily measured or rewarded by policy interventions.

This paper adds to a growing literature on the long term effects of education programs. Earlier studies focused on the long-term effects of compulsory schooling laws on adult educational attainment (Angrist and Krueger, 1991) and on adult health (Lleras-Muney, 2005), for example. More recent studies have addressed school programs aimed at improving the quality of education, in addition to increasing attainment. Most of these studies evaluated standardized test scores as an effective measure of success. However, an equally relevant question is the extent to which educational interventions lead to long-term improvements in well-being – measures assessed by attainment in life.

Puzzling and conflicting results from several evaluations make this a highly salient issue. Three small-scale, intensive preschool experiments produced large effects on contemporaneous test scores that quickly faded (Schweinhart et al., 2005; Anderson, 2008). Non-experimental evaluations of Head Start, a preschool program for poor children, revealed a similar pattern, with test-score effects dissipating by middle school. However, in each of these studies treatment effects re-emerged in adulthood, in the form of increased educational attainment, enhanced labor market attachment, and reduced crime (Deming, 2009; Garces et al., 2002; Ludwig and Miller, 2007). Other studies have shown evidence for the effect of investments in childhood on post-secondary attainment (Krueger and Whitmore 2001, Dynarski et al 2011). Recently, Chetty et al. (2011 and 2014) examined the longer-term effect of value-added measures of teacher quality in a large urban school district in the United States. They reported significant effects on earnings at age 27, even though the effect on test scores had faded away much earlier. Dustmann et al (2012), however, found that attending a better school in Germany had no effect on tertiary school attainment or labor-market outcomes. Even though the ultimate goal of education is to improve lifetime well-being and there is much uncertainty about the long term gains from such programs, there have not been any studies that focus on the long term effect of teachers' PFP. Determining which interventions are more effective in improving long-term outcomes is critical for refining the effectiveness of education and school resource allocation.

The remainder of this paper is as follows. Section 2 describes the PFP experiment and Section 3 describes the data. Section 4 outlines the identification and econometric model and Section 5 presents the empirical findings. Section 6 concludes.

2. The Pay for Performance Experiment

The popularity of teacher incentive programs is increasing. Performance-related pay for teachers is being introduced in many countries, amidst controversy and opposition from teachers and unions. The rationale for these programs is that incentive pay may motivate teachers to improve their performance (Lazear 2000 and 2001, Lavy 2002, 2007 and 2009, Neal 2011, Duflo et al. 2012). Opponents of teachers' incentive programs argue that schools may respond to test score-based incentives in perverse ways, such as by cheating in grading and teaching to the test (Glewwe et al, 2010, Neal 2011), leading to short term gains in performance but not to the long-term accumulation of human capital. This paper presents evidence on a wide array of lifetime outcomes.

2a. Secondary Schooling in Israel

High school students in Israel sit for the *Bagrut* examinations, a set of national exams in core and elective subjects that are administered from during $10^{th} - 12^{th}$ grade. This testing program is similar to that found in many countries in Europe and elsewhere. The final matriculation score in a given subject is the mean of two intermediate scores. The first is based on the score in the national exams, which are "external" to the school because they are written, administered, supervised and graded by an independent agency. Scoring for these exams is anonymous: the external examiner is not told the student's name, school or teacher. Exams are held in June and January, and all pupils are tested in a given subject on the same date. The national exams are graded centrally by two independent external examiners and the final external score is the average of the two. The second intermediate score is based on a school-level ("internal") exam that mimics the national exam in material and format but is scored by the student's own teacher.

Some subjects are mandatory and many must be taken at the level of three credits as a minimum. Subjects that award more credits are more difficult. English and math are among the core compulsory subjects and must be studied at one of three levels: basic (3 credits), intermediate (4 credits) and advanced (5 credits). A minimum of twenty credits is required to qualify for a matriculation certificate. About 45 per cent of high-school seniors received matriculation certificates in 1999 and 2000, i.e., passed enough exams to be awarded twenty credits and satisfied the distributional requirement by the time they graduated from high school or shortly thereafter (Israel Ministry of Education, 2001). The high school matriculation certificate in Israel is a prerequisite for university admission and is one of the most economically important education milestones.

2b. The Teacher-Incentive Experiment

In early December 2000, the Ministry of Education unveiled a new teachers' bonus experiment in forty-nine Israeli high schools. The main feature of the program was an individual performance bonus paid to teachers on the basis of their own students' achievements. The experiment included all English, Hebrew, Arabic, and mathematics teachers in these schools 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 activity for principals and administrators of the forty-nine schools. The program was described to them as a voluntary three-year experiment.[3] All the principals reacted enthusiastically except for one, who decided not to participate in the program.[4]

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

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

Each teacher entered the tournament as many times as the number of classes he/she taught and was ranked each time on the basis of the mean performance of each of his/her classes. The ranking was based on the difference between the actual outcome and a value predicted on the basis of a regression that controlled for the students' socioeconomic characteristics, the level of their study program in the relevant subject (basic, intermediate and advance), grade (10th, 11th and 12th), grade size, and a fixed school-level effect.[5] The school fixed effects imply that the predicted values were based on within school variation among teachers, and the teachers were told explicitly that they were to be compared to other teachers of the same subject in the same school. Separate regressions were used to compute the predicted pass rate and mean score, and each teacher was ranked twice – once for each outcome – using the size of the residual from the regressions. The school submitted student enrollment lists that were itemized by grades, subjects, and teachers. The reference population was those enrolled on January 1st, 2001, the starting date of the program. All students who appeared on these lists but did not take the exam (irrespective of the reason) were assigned an exam score of zero.

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

Three formal rules guided the assignment of schools to the program: only comprehensive high schools (comprising grades 7–12) were eligible, the schools had to have a recent history of relatively poor performance in the mathematics or English matriculation exams,[7] and the most recent school-level matriculation rate must be equal to or lower than the sector-specific national mean (for the Jewish secular sector this rate is 45 per cent and for Jewish religious and Arab sectors this rate is 43 percent).

[5] Note that the regression used for prediction did not include lagged scores, so that teachers would have no incentive to play the system, for example by encouraging students not to do their best in earlier exams that do not count in the tournament. This feature would have been important had the program continued.

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

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

106 schools met the first two criteria but seven of them were disqualified because they were already part of other remedial education programs. Therefore, there were 99 eligible schools of which 49 met the third criterion. However, as noted above, one school declined to participate in the program and thus the actual number of participants was 48 schools (treated sample).[8]

It was not possible for a school to manipulate the assignment variable in order to be included in the program. First, the schools only learned about the program when the Ministry informed the head of the school that the school had been selected to the program, and invited heads and other school staff to the orientation meeting at the Ministry of Education. At that meeting, the principal investigator ((the designer of the program) explained the details of the program and time schedule. Second, the *Bagrut* Exams Division at the Ministry computed the temporary *Bagrut* rate for each school (the variable that we use as the running variable for inclusion in the program) in October 2000, well before the schools were informed about their selection to the program. The temporary *Bagrut* rate is computed for internal Ministry purposes and is not communicated to schools. Schools learn about their final *Bagrut* rate at the very end of the marking process, where *Bagrut* eligibility is determined for all students. Third, using the national average *Bagrut* rate as the threshold for eligibility to the program was the principal investigator's decision and therefore schools had no way of knowing about it. Finally, note that schools could not have manipulated their 'true' final *Bagrut* rate either, because it is simply the school mean matriculation rate computed by the Ministry of Education.

The program included 629 teachers, of which 207 competed in English, 237 in mathematics, 148 in Hebrew or Arabic, and 37 in other subjects that schools chose instead of Hebrew. 302 teachers won awards—94 English teachers, 124 math teachers, 67 Hebrew and Arabic teachers, and 17 among the other subjects. Three English teachers won two awards each, twelve math teachers won two awards each, and one Hebrew teacher won two first-place awards totaling $15,000.

We did a follow-up survey of teachers in the program in the summer vacation following the end of the school year. Seventy-four per cent of teachers were interviewed. Very few of the intended interviewees were not interviewed. Failure to be interviewed was mostly due to incorrect telephone numbers or teachers who could not be reached by telephone after several attempts. The survey results show that 92 per cent of the teachers knew about the program, 80 per cent had been briefed about its details—almost all by their principals and the program coordinator—and 75 per cent thought that the information was complete and satisfactory. Almost 70 per cent of the teachers were familiar with the award criteria and about 60 per cent of them thought they would be among the award winners. Only 30 per cent did not believe they would win; the rest were certain about winning. Two-thirds of the teachers thought that the incentive program would lead to an improvement in students' achievements.

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

3. Data

In this study, I use data from the administrative files of the participants in the treatment and control groups. The students in the sample graduated from high school between 2000 and 2001, and in 2013 they are adults aged 30-31. I use several panel datasets from Israel's National Insurance Institute (NII). The NII is responsible for social security and mandatory health insurance in Israel. NII allows restricted access to this data in their protected research lab. The underlying data sources include: (1) the population registry data, which contains information on marital status, number of children and their birth dates; (2) NII records of tertiary education enrollment from 2000 through 2013, based on annual reports submitted to NII every fall term by all of Israel's tertiary education institutions. Based on this annual enrollment data I computed the number of years of tertiary schooling[9]; (3) Israel Tax Authority information on income and earnings of employees and self-employed individuals for each year during 2000-2012. This file includes information both for the students and their fathers and mothers; (4) NII records on unemployment benefits for the period 2009-2012, and marriage and fertility information as of 2012. The NII linked this data to students' background data that I used in Lavy (2009). This information comes from administrative records of the Ministry of Education on the universe of Israeli primary schools during the 1997-2002 school years. In addition to an individual identifier, and a school and class identifier, it also included the following family-background variables: parental schooling, number of siblings, country of birth, date of immigration if born outside of Israel, ethnicity and a variety of high school and high school achievement measures. This file also included a treatment indicator, and cohort of study.

The NII data track all individuals in Israel and also those who leave the country providing they continue to pay National Insurance Tax. The basic sample of students from all 98 schools that were eligible to participate in the program included 25,588 students. Only 153 students from this sample (0.6 percent) were not found in the Population Registry at NII, 68 were from the control group (out of 12,500) and 85 from the treated group (out of 13,068). The proportion in the natural experiment and in the RD sample are very similar to the proportion in the eligible sample. This means our long term analysis tracks 99.4 percent of the students to adulthood.

3a. Post High School Academic Schooling in Israel

Israel has seven universities (one of which confers only Master and PhD degrees), and more than 50 academic colleges that confer undergraduate degrees (some of these also give masters degrees).[10] The universities require a *bagrut* diploma for enrollment. Most academic colleges also

[9] The NII, which is responsible for the mandatory health insurance tax in Israel, tracks post-secondary enrollment because students pay a lower health insurance tax rate. Post-secondary schools are therefore required to send a list of enrolled students to the NII every year. For the purposes of this project, the NII Research and Planning Division constructed an extract containing the 2001–2013 enrollment status of students in our study.

[10] A 1991 reform sharply increased the supply of post-secondary education in Israel by creating publicly funded regional and professional colleges.

require a *bagrut*, although some accept specific *bagrut* diploma components without requiring full certification. For a given field of study, it is typically more difficult to be admitted to a university than to a college. The national university enrollment rate for the cohort of graduating seniors in 1995 (through 2003) was 27.6 per cent and the rate for academic colleges was 8.5 percent.[11]

The post-high school outcome variables of interest in this study are indicators of ever having enrolled in a university or in an academic college as of the 2013 school year, and the number of years of schooling completed in these two types of academic institutions by this date. I measure these two outcomes for the 2000 and 2001 12th grade students. The year refers to spring of the year the students were expected to graduate. Thus the 2000 grade 12 cohort is the control group (C) and the 2001 grade 12 cohort is the treatment group (T). Even after accounting for compulsory military service[12], I expect most students who enrolled in academic post-high school education, including those who undertook post-graduate studies, to have graduated by the 2013 academic year.

3b. Definitions of Outcomes in Adulthood

***Post-Secondary Academic Schooling*:** In the NII data, I observe two sets of post-secondary outcomes for the students in the sample. First, I observe year-by-year enrollment in post-secondary education, including the type of institution attended, if any. Therefore, the first measure is cumulative ever enrollment by type of institution and by year. For example, the ever enrollment measure for university 12 years after high school completion is an indicator of ever being enrolled in university 12 years after finishing high school. The second measure is the number of years of education completed in each type of institution by a given year.

***Labor Market Outcomes*:** I observe year-by-year labor market outcomes from high school graduation to 2012, including employment status, months of work during the year, and annual earnings. Individual earnings data comes from the Israel Tax Authority (ITA). Only individuals with non-zero self-employment income are required to file tax returns in Israel, but ITA has information on annual gross earnings from salaried and non-salaried employment and transfers this information annually to NII, including the number of months of work in a given year. NII produces an annual series of total annual earnings from salaried and self-employment. Following NII practice, individuals with positive (non-zero) number of months of work and zero or missing value for earnings are assigned zero earnings. 14.1% of individuals (students) have zero earnings at age 30-31 in our basic sample and 16.6% have zero earnings in this sample. To account for earnings data outliers, I dropped from the sample all observations that are six or more standard deviations away from the mean. Very few observations are dropped from the sample in each of the years and the results are not qualitatively affected by this sample

[11] This data is from the Israel Central Bureau of Statistics, Report on Post-Secondary Schooling of High School Graduates in 1989–1995 (available at: http://www.cbs.gov.il/publications/h_education02/h_education_h.htm).

selection procedure. To account for age differences of the different cohorts included in the sample, the outcomes are adjusted for years since graduating high school. The same earnings data is also available for the parents of the students in our sample, for 2000-2002 and 2008-2012. I compute the average earnings of each parent and of the household for 2000-2002 and use it as an additional control in a robustness check of the evidence presented in this paper. These data were not available for the analysis of the effect of the program on short-term outcomes. I also use as additional outcomes the NII indicator of being *Eligible for Unemployment Benefit* and the annual amount of *Unemployment Benefit Compensation*.

***Personal Status Outcomes*:** The Population Registry is only available to us for 2012. However, the dates of each marriage and birth event are reported in the data and therefore I can adjust the demographic outcomes for years since graduating high school. These outcomes include indicators for *Marriage Status* and for *Having Children.*

4. Identification

4a. Measurement Error in the Assignment Variable

There is a measurement error in the assignment variable of eligible schools for the program, which generates a natural experiment, given the true value of the assignment variable. This allows for identification of the causal effects of the program, by comparing school that were randomly but erroneously assigned to treatment, to those that should have been assigned. The rules of the program limited assignment to schools with a 1999 matriculation rate equal to or lower than 45 per cent for Jewish secular schools and 43 per cent for religious and Arab schools. However, the matriculation rate used for assignment was an inaccurate measure of this variable. The data given to administrators were culled from a preliminary and incomplete file of matriculation status. For many students, matriculation status was erroneous, as it was based on missing or incorrect information. The Ministry later corrected this preliminary file, as it does every year. There are many requirements to complete the matriculation process that tend to vary by school type and level of proficiency in each subject. The verification of information between administration and schools is a lengthy process. The first version of the matriculation data becomes available in October and is finalized in December. As a result, the matriculation rates used for assignment to the program were inaccurate in a majority of schools.[13] Figure

[13] Several examples demonstrate the sources of the measurement error in the preliminary school mean matriculation rate. Most of the following examples are related to delays in transmitting the exam grades of some students to the Ministry of Education, so that the school mean is computed based on an incomplete and potentially non-representative sample. This naturally results in a gap between the preliminary (measured with error) and the final (correct) school mean matriculation rate. For example, immigrant students can ask to write exams in their native language; their papers are graded by special examiners, which can take longer and lead to delays in final grades reaching the Ministry of Education. However, as shown in Lavy (2009) and presented also later in this paper online appendix table, tests of balance between treatment and control schools show no differences in proportion of immigrant students. Another example is suspected cheating, requiring sometime re-grading of the papers of all students who took the test in the same room. Some students are tested verbally and their grade is communicated to the Ministry separately, which often takes longer. High achieving students can write a thesis in

A1 in the online appendix (Figure 1 in Lavy 2009) presents the relationship between the correct matriculation rates and those erroneously measured for a sample of 507 high schools in Israel in 1999. Most (80 percent) of the measurement errors were negative, 17 per cent were positive, and the rest were free of error. The deviations from the 45-degree line do not seem to correlate with the correct matriculation rate. This may be seen more clearly in Figure A2 in the online appendix (Figure 2 in Lavy 2009), which demonstrates that the measurement error and the matriculation rate do not co-vary. However, if a few extreme values (five schools) are excluded, the correlation coefficient is effectively zero. Although the figure possibly suggests that the variance of the measurement error is lower at low matriculation rates, this is most likely due to the floor effect that bounds the size of the negative errors: the lower the matriculation rate, the lower the absolute maximum size of the negative errors. Similar results are observed when the sample is limited to schools with a matriculation rate higher than 40 percent. In this sample, the problem of the bound imposed on the size of the measurement error at schools with low matriculation rates is eliminated. I also examined a sample that was limited to the eligible schools, the results of which are identical to those in Figures A1 and A2.

There can be a further check on the randomness of the measurement error based on its statistical association with other student or school characteristics. Table A1 (Table 2 in Lavy 2009) presents the estimated coefficients from regressions of the measurement error on student characteristics, lagged students' outcomes, and school characteristics of the 2001 (treatment) high school seniors. Each entry in the table is based on a separate regression which was run with school-level means of all variables, separately for the whole sample and for the eligible sample. There are 12 estimates for each sample, and only a few are significantly different from zero. Furthermore, the variables that are significant are different across samples, suggesting that these are transitory and random differences. The results presented in Figures A1 and A2 and online appendix Table A1 show no evidence of a significant association between the measurement error in 1999 and the observable characteristics, therefore the likelihood that the measurement error is correlated with other unobserved confounders is also very low. Admittedly, however, it is not possible to test the assumption that the measurement error is not correlated, through some unobservables, with the change in outcomes from 2000 to 2001.

Identification based on the random measurement error can be presented formally as follows:

Let $S = S^* + \varepsilon$ be the error-affected 1999 matriculation rate used for the assignment, where S^* represents the correct 1999 matriculation rate and ε the measurement error. T denotes participation status, with $T = 1$ for participants and $T = 0$ for non-participants. Since $T(S) = T(S^* + \varepsilon)$, once I control for S^*,

addition to a matriculation exam in a given subject. Often the final grades of these students are processed at a later date. Given that in these examples there is no obvious average selection it is not surprising that the measurement error in the school matriculation rate is not correlated with the true matriculation rate nor with any observable school and student characteristics, as I show below.

assignment to treatment is random ("random assignment" to treatment, conditional on the true value of the matriculation rate). The presence of a measurement error creates a natural experiment, where treatment is assigned randomly, conditional on S^*, in a sub-sample of the 98 eligible schools. Eighteen of the eligible schools had a correct 1999 matriculation rate above the threshold. Thus, these schools were "erroneously" chosen for the program. For each of these schools, there is a school with an identical correct matriculation rate but with a draw from the (random) measurement error distribution which is not large (and negative) enough to drop it below the assignment threshold. Such pairing of schools amounts to non-parametrically matching schools on the basis of the value of S^* (see Figure A3 in the online appendix for a graphical presentation of this matching which is reproduced from Lavy 2009 Figure 3). Therefore, the eighteen untreated schools may be used as a control group for identification of the effect of the program. Since some of the control schools are matched to more than one treated school, and vice versa, weighted regressions are used to account for the treatment-control differences in sample size within the matched groups (the weights are presented in Table A4 in the online appendix of Lavy 2009).

To further alleviate concerns about the possibility of manipulation of the matriculation rates by schools or by Ministry officials who were aware of the program, I first show histograms of schools' density of the matriculation rate, based on the sample of 98 eligible schools, which show that there is no bunching at the thresholds of assignment to the program. Secondly, I present results of McCrary's (2008) density discontinuity tests, here also based on the sample of 98 eligible schools, which formally and empirically rule out evidence of manipulation at eligibility thresholds. Online appendix Figures A4 and A5 present separate histograms for the Jewish secular schools for which the matriculation eligibility rate is 45 and for Jewish religious schools and Arab schools for which the matriculation eligibility rate is 43. First note that at each rate there are only a few schools, 1, 2 or 3. Secondly, the two figures below clearly show no evidence of bunching at the relevant thresholds. In the first figure for Jewish secular schools there are 3 schools at 45 but also 3 schools at 43 and there are 2 schools in each of 5 other rates. Similarly, in the second figure there are 3 schools at 43 but also 3 schools at 41, 38, and 31 and many rates have two schools. Figure A6 in online appendix presents the McCrary (2008) density discontinuity tests that statistically illustrate that there is no discontinuity of the matriculation rate at the treatment cutoffs (45 and 43) and at nearby rates (44, and 42). The graphs and tests are clear evidence of no discontinuities in the density at the eligibility cutoffs.

4b. Regression Discontinuity

To check the robustness of the results based on the natural experiment sample, I use an additional alternative method, an RD design like sample in a difference in differences setup, that provides additional supporting results regarding the causal effect of the pay experiment. Given that the rule governing selection to the program was simply based on a discontinuous function of a school

observable, the probability of receiving treatment changes discontinuously as a function of this observable. The discontinuity in our case is a sharp decrease (to zero) in the probability of treatment beyond a 45 per cent school matriculation rate (which is the 'running' variable in this one threshold RD design) for Jewish secular schools and beyond 43 per cent for Jewish religious and Arab schools. I exploit this sharp discontinuity to define a treatment sample that included schools that were just below the threshold of selection to the program and a comparison group that included untreated schools that were just above this threshold. The time series on school matriculation rates show that the rates fluctuate from year to year for reasons that transcend trends or changes in the composition of the student body. Some of these fluctuations are random. Therefore, marginal (in terms of distance from the threshold) participants may be similar to marginal nonparticipants. The degree of similarity depends on the width of the band around the threshold. Sample size considerations exclude the possibility of a bandwidth lower than 10 percent, and a wider band implies fluctuations of a magnitude that are not likely to be related to random changes. Therefore, a bandwidth of about 10-12 per cent seems to be a reasonable choice in our case. The main drawback of this approach is that it produces an estimate from marginally (relative to the threshold) exposed schools only. However, this sample may be of particular interest because the threshold schools could be representative of the schools that such programs are most likely to target.

There are 13 untreated schools with matriculation rates in the 0.46–0.52 range and 14 treated schools in the 0.40–0.45 range. The 0.40–0.52 range may be too large, but I can control for the value of the assignment variable (the mean matriculation rate) in the analysis. I also present evidence based on two wider bands, which allow assessment of the sensitivity of the estimates to the width of the band.[14]

4c. Descriptive Statistics and Tests for Balance Between Treatment and Control Groups

The treatment and control groups in the natural experiment and in the RD samples are perfectly balanced in student and school characteristics (see Lavy 2009 Table 3 and Table 6). Table 1 presents detailed summary descriptive statistics and treatment-control balancing tests for the long term outcome variables for 2012, by treatment and control group, for the pre-program cohort that graduated high school in 2000. Columns 1-3 present this evidence for the natural experiment sample and columns 4-6 for the RD sample. The results presented in this table show treatment-control similarities in post-secondary enrollment (panel A) and completed years of schooling (panel B) of the pre-program cohort for both samples. For example, in the natural experiment sample, the mean years of academic college is 0.379 in the treatment group and 0.365 in the control group; the difference, 0.014, is not statistically different from zero. Similar treatment-control balancing is seen in panel D in marriage and fertility

[14] Note, also, that there is some overlap between this sample and the natural experiment sample. Eleven of the 14 treated schools and 8 of the 13 control schools in the RD sample are also part of the natural experiment sample, leaving only six schools (3 control and 3 treated), which are included in the former but not in the latter. However, there are 17 schools in the natural experiment sample (7 treated and 10 control) that are not included in the RD sample, which suggests that there is enough "informational value added" in each of the samples.

outcomes and in pane E in parental average earnings in 2000-2002, the years that students in the sample were in high school. Note that this information became available only recently through the NII data, and I, therefore, add it now to the treatment-control balancing analysis.

The balancing tests for the pre-program cohort labor market outcomes in 2011 (panel C) reveal a different pattern. In particular, the treatment-control difference in mean earnings in both samples is large and significant. For example, in the natural experiment sample, the average annual earnings of the treatment group is 62,001 NIS, equivalent to $16,318 (based on the prevailing exchange rate of 3.8 Israeli Shekels to one US Dollar), and of the control group it is 67,682 NIS ($17,813), and the difference is marginally statistically different from zero (-5,681, se = 3,265). However, this imbalance is fully explained in both samples by an imbalance in number of Arab schools in the treatment and control group, as there are three such schools in the treatment group while there are no Arab schools in the control group. This imbalance translates to pre-program labor market gaps between treatment and control because of considerable differences in employment and earnings between Arabs and Jews in Israel. These differences are well documented and mostly explained by differences in schooling, norms and culture and discrimination.[15] For example, the differences in wages between Arab and Jewish workers are significant because of the very low labor force participation of Arab women relative to Jewish women. Such gaps are observed in both of our samples. For example, in the natural experiment pre-program 2000 cohort sample, the employment rate in 2012 is lower for Arabs by 16.8 percent, the average annual months of work is lower by 2.5 months and the earnings of Arabs is lower by NIS 26,846 (almost 40 percent difference). Clearly, the means of all labor market outcomes are significantly lower in the sample of Arabs that are included in the treated group. Note also that the means in the sample of Jewish schools that are part of the treatment group is almost identical to the means of the control group schools, which as noted above includes only Jewish schools. It implies that the labor market outcomes imbalances shown in Table 1, panel C, are all due to the students in Arab schools that are part of the treatment group. These imbalances are eliminated once I add to the balancing regressions a control for the type of school (Arab or Jewish). These results are presented in online appendix Table A2, which reports balancing tests with and without control for type of school for earnings and months of work for the 2000 (pre-treatment) cohort, and a comparison of means of these labor market outcomes for the 2001 (treatment) cohort. Panel A presents the results based on the natural experiment sample. The pre-program treatment-control raw earnings gap is -5,681 and it falls to -601 when a control for an Arab school is added to the balancing regression. This difference is not statistically different from zero. An opposite pattern is seen for the labor market outcomes of the 2001 (treatment) cohort: negative or small positive treatment-control differences that become positive and large once the differences are

[15] See Yashiv and Assali (2006), "Why Do Arabs Earn Less than Jews in Israel?", Tel Aviv University.

conditioned on school type. A similar pattern is seen in panel B that report the evidence for the RD sample.

4d. Estimation Model

The following model is used as the basis for regression estimates using the natural experiment sample:

$$Y_{ijt} = \alpha + X_{ijt}'\beta + Z_{jt}'\gamma + \delta T_{jt} + \Phi_j + \eta D_t + \varepsilon_{ijt} \quad (1)$$

Where, *i* indexes individuals; *j* indexes schools; t indexes the cohort in years 2000 or 2001 where t=2001 indicates those students who were in the graduating cohort (grade 12th) in the 2000-01 school year, and who therefore had teachers who received incentive pay based on spring 2001 test scores, and t=2000 indicates students who were in the graduating cohort (grade 12th) in the 1999-2000 school year, the year before the one-year incentive experiment was conducted. *T* is the assigned treatment status. *X* and *Z* are vectors of individual and school level covariates, and D_t denotes year effects with a factor loading η. *X* includes number of siblings, gender dummy, father's and mother's education, a dummy indicator for immigration status, a dummy variable indicating Asian/African ethnicity, student's lagged achievements before treatment (number of credits gained in the relevant subject (i.e., math or English), the number of credit units attempted, the average score in those attempted units, overall credit units awarded). *Z* includes the school mean one- and two-year lagged matriculation rate.

The treatment indicator T_{jt} is equal to the interaction between a dummy for treated schools and a dummy for the year 2001. The regressions will be estimated using pooled data from both years (the two adjacent cohorts of 2000 (pre-treatment) and 2001 (treatment), stacked as school panel data with fixed school-level effects (Φ_j) included in the regression. The resulting estimates can be interpreted as an individual-weighted difference-in-differences procedure comparing treatment effects across years. The estimates are weighted by the number of students in each school. The introduction of school fixed effects controls for time-invariant omitted variables. Standard errors are clustered at the school level, allowing for correlated errors within a cluster. The model specified in equation (1) is also used for estimation of treatment effect based on the RD sample.

5. Empirical Evidence

5a. Effect on Post-Secondary Education Attainment

The program had positive and significant short term effects on high school English and math outcomes at the end of high school (Lavy 2009). Since the program increased exam participation, the average score, and the passing rate in the math and English matriculation exams, one should expect also a positive effect on the overall summary outcomes of the matriculation exams. In the natural experiment sample, for example, the PFP program led to an increase of the average matriculation score by 2.8 points (se=0.892), and to an increase in the matriculation rate by 3.6 percentage points, which amounts to an 8 per cent improvement. The average number of credit units increased by 0.8, the number of credits in science increased by 0.6 units (a 25% per cent increase), and the number of subjects studied at the most

advanced level (5 credits) increased by 0.1. These results are presented in online appendix Table A3. Improvement in these summary achievement measures should lead to an increase in post-secondary academic schooling because they are used as admission criteria for various academic institutions and study programs.

I first show a graphical representation of the effect of the program on post-secondary education, focusing on the two branches of academic post-secondary education in Israel. The first includes the seven research universities in Israel that confer BA, MA and PhD degrees. Admission to these schools requires a matriculation diploma, with an intermediate or advanced level in English (that is, the level of proficiency in English required for admission to university is higher than the basic level that is required for a matriculation diploma) and at least one matriculation subject at an advanced level. Nationwide, about 35 per cent of all students are enrolled in one of the seven universities. The second branch includes more than 50 academic colleges that mostly confer a BA degree, and generally offer social sciences, business and law degrees.

In Figure 1, I measure the treatment effect for each year after high school graduation and trace the dynamic pattern for university enrollment for the natural experiment sample. To do so, I run a separate difference in differences regression (equation 1) for each year since high school graduation. I then plot the coefficients of these regressions around a 90 per cent confidence interval based on standard errors that are clustered at the school level. Note that both the ever-enrolled variable and the years of schooling are cumulative variables. Hence, I expected the effects to be either flat or increasing over time. This treatment effect becomes positive from year three after graduating from high school and it reaches its height, at five percentage points, from the eighth year after high school graduation, remaining flat afterwards.[16] This pattern likely reflects the fact that high school graduates who do not enroll in post-secondary education in the first eight years are very unlikely to continue their education later in life. In contrast, the effect on years of education accumulates over time (Figure 1A). Although most of the increase happens in the first eight years, the effect seems to be increasing even after 12 years after graduation, reaching a peak of 0.25 years. The fact that the increase keeps accumulating even 12 years after high school graduation suggests that focusing on outcomes immediately after graduation may underestimate the long-term effects. Note that the effect on the intensive margin seems to operate beyond the increase in enrollment. Given a five percentage-point increase in enrollment and a typical duration of 3-4 years, I would expect education to increase by only 0.15-0.20 years. The fact that the effect on years of education is larger than 0.20 years suggests that the program induced treated students to stay longer and complete longer programs.

The effect size on enrollment and years of education can be compared to the mean enrollment rate for the treated group, which increases gradually from year one and is highest at 20 per cent thirteen

[16] The emergence of the treatment effect after three years is reasonable given that most of the female students are in military service for two years following high school graduation and for boys this period is three years.

years later. The mean of university years of schooling in the treatment group is 0.8.[17] Figure 2 and Figure 2A present the estimated effects on academic college outcomes and here the estimated effect is negative and small, practically close to zero. The same figures based on the RD sample are identical to those obtained based on the natural experiment sample.

Table 2 presents the controlled (panel A) and uncontrolled (panel B) DID estimates and their standard errors for the impact of the PFP experiment on university and academic college education, measured at the end of the period of study. The table presents results based on the natural experiment sample (columns 1-4) and the RD sample (columns 5-8). Effects on enrollment are presented in columns 1-2 and 5-6. Evidence for years of schooling is presented in columns 3-4 and 7-8.

Discussing first the controlled DID results, enrollment in university increased by 4.8 percentage points and this effect is precisely measured (se=0.019). This gain, relative to the pre-program mean of students in treatment schools (21.6 percent), is a 22 per cent increase. This increase in enrolment led to a 0.250 increase in completed years of university education, reflecting a 30 per cent increase relative to the baseline mean of 0.825 years of university education. The relative gain in university enrollment and completed years of education have similar magnitude, both being large relative to some other education interventions or policy changes, for example, compared to the gains from an increase in compulsory schooling. The effect on academic college enrollment and years of education is negative but close to zero and not precisely measured. This pattern may suggest some compositional change in post-secondary education but the offsetting decline in academic college attendance is too small to be economically meaningful. The differences between treatment effect on university and on academic college outcomes are statistically different from zero.[18] In the third row of Table 2, I present the estimated effect on all post-secondary education. The program effect on overall enrollment and years of post-secondary education is lower than the effect on university education, reflecting its negative effect on college enrollment and years of education. For example, the effect on post-secondary education based on the natural experiment sample is 0.170 years (se=0.089), lower than the effect on university education, 0.25 years. The difference between the two is almost exactly the negative effect on college years of education, -0.072. The conclusion is that the program induced some substitution towards enrollment in universities and away from academic colleges, but also had a significant effect on the extensive margin of university education.[19]

[17] The yearly university enrollment rate is highest at year 4-7 and then starts to decline until practically leveling at close to zero at year 13 (this result is not shown in the paper due to space limitations).

[18] The P values for these differences across samples (natural experiment and RD) fall in the range 0.0533-0.0004, indicating that they are all significantly different from zero. These values are not presented in Table 2 so as not to overload it with figures.

[19] The identification based on the natural experiment predicts that conditional on the true matriculation rate, a simple difference between treated and control schools should yield estimates close to the natural experiment sample results. Indeed, estimates of the 2001 simple differences between treated and control schools, conditional only on the true matriculation rate, are very similar to the treatment estimates obtained from the DID model. These results are available from the author.

The small number of clusters (36 schools) may bias downward the clustered standard errors estimates. To make sure that this is not the case I use an alternative procedure to compute standard errors, the wild bootstrap procedure which is less sensitive to the number of clusters. In online appendix Table A4 I reproduce the estimates presented in Table 2 while adding the bootstrap standard errors along with the school level clustered standard errors. The wild bootstrap standard errors are marginally higher than the school level clustered standard errors but overall all estimates in Table 2 that are statistically different from zero based on the school level clustering are also statistically significant and in most cases at the same level of significance when using the wild bootstrap procedure. For example, the two standard errors for the effect on university enrollment (0.048) are 0.019 (clustered) and 0.024 (bootstrap).

The estimates based on the RD sample are presented in columns 5-8 of Table 2. They are very similar to the estimates based on the natural experiment sample.[20] The level of statistical significance of all estimates is not changed when using the bootstrap standard errors (see online appendix Table A4).

I also re-estimated this treatment effect with broader bandwidth, approximately 0.38-0.53[21] and 0.37-0.54, instead of the 0.40-0.52 original bandwidth). As can be seen in online appendix Table A5, the results are not sensitive to the small changes in the bandwidth. For example, based on the bandwidth of 0.38-0.53, the effect on completed university years of schooling is 0.221 years and based on the bandwidth of 0.37-0.54 the respective estimated effect is 0.219.

In panel B of Table 2, I present uncontrolled simple differences in differences estimates, using only a post-treatment dummy, a treated school dummy interacted with the post dummy and school fixed effects. The estimated treatment effects based on this limited control specification are positive and significant for all university attainment outcomes, both in the natural experiment and the RD sample. These estimates are very close to those of estimates based on the full specification presented in panel A, confirming what is expected given the balancing tests presented in Table 1. For example, the simple difference in differences estimate on university years of schooling is 0.206 (0.102) and the respective controlled difference in difference estimate is 0.250 (0.094).

In Table 3, I present cross-section regressions separately for the pre-treatment cohorts that graduated high school in 1999 and 2000, and for the post-treatment cohort that graduated high school in 2002, along with a similar regression for the 2001 (treatment) cohort. Since data for additional earlier untreated cohorts is not available to us, the evidence in this table provides alternative data for examining the parallel trends assumption. The estimates presented are based on a specification that includes the same controls that are included in the DID regressions, including controls for students' and schools' characteristics. The 2001 cohort treatment-control differences in the university ever enrollment and

[20] The difference in differences regressions using the RD sample include the same control variables as the natural experiment sample difference in differences regressions. Note that in these regressions the running variable is absorbed by the school fixed effect.

[21] Note that this range starts at 38 because there are no treated schools at 39 percent.

years of schooling are positive and they stand out relative to very similar differences in the other three years. Considering, for example, the treatment-control difference in university years of schooling in 1999 and 2000 (pre-treatment) cohorts, and 2002 (post-treatment) are -0.084, 0.058, and 0.040, respectively, and all three have a very similar standard error. The treatment-control difference in 2001 (treatment) cohort is 0.254. This comparison shows that the DID estimated effect on university schooling of the 2001 (treatment) cohort is mostly due to an improvement in outcomes of the treated group in the 2001 cohort.

5b. Using Younger and Older Untreated Cohorts for Placebo Control Experiments

I assessed the robustness of the estimates presented in Table 2 in two ways. First, I estimated the treatment effect using earlier or later cohorts as a non-treated period and using the same samples of schools. In online appendix Table A6, panel A, I present estimates where the 1999 12th grade cohort is used as the 'before' untreated cohort versus the treated 12th graders of 2000. The estimated effects on post-secondary outcomes are small and not significantly different from zero. For example, the effect on university enrollment is 0.017 (se=0.019) and on university years of schooling is 0.024 (se=0.075). In panel B I present estimates based on the 12th grade cohort of 2000 as the untreated cohort versus the untreated 12th graders of 2002.[22] For example, the effect on university enrollment is 0.015 (se=0.022) and on university years of schooling is 0.083 (se=0.079). Here again, all estimates are small and not significantly different from zero.

In a second robustness check, I assigned the treatment status randomly to schools. For example, the effect on university enrollment is -0.043 (se=0.030) and on university years of schooling is -0.142 (se=0.154). These placebo effects on schooling outcomes are actually negative though not different from zero, and importantly they are significantly different from the treatment effect estimates presented in Table 2. These results are presented in online appendix Table A7.

5c. Effect on Employment and Earnings

We expect the increase in the quality and the quantity of high school and post-secondary education to result in better labor market outcomes in adulthood. In Figures 3-4, I repeat the year-by-year analysis, focusing on labor market outcomes based on the natural experiment sample. The figures show the estimated effects by years since graduation from high school. The employment and earnings data are available until year 2012, so eleven years is the longest period since graduating high school for which I examine the effect of the program. Overall, I find an increasing pattern in both employment and earnings. The effects become significantly different from zero about 9-11 years after high school graduation. Perhaps reflecting the higher enrollment in post-secondary education, the effect on employment is initially negative and increases thereafter. The effects on earnings follow a similar

[22] We can view the 12th grade students in 2002 as untreated because most of them did not take a math or English matriculation exam in the previous year (when the experiment took place), as 11th grade students in 2001.

pattern. As treated students spend more years on average in the education system and appear on average to start working later, I expect the effect on earnings to be initially negative and to increase as students complete their post-secondary education and accumulate labor market experience. Indeed, I find that the effects are initially negative and become significantly different from zero by the end of our sample period. In the following paragraph I describe these results in more detail.

Figures 3 presents the yearly estimates for employment. The large positive effect on employment in years 0-2 is likely a result of the treatment group having 15 percent of its students from Arab schools while the control schools have no such schools. Since Arab students are not drafted to military service (they can volunteer but very few do), they join the labor market or enroll in post-secondary schooling soon after graduating high school. After completion of compulsory military service of the Jewish students in the sample (2 years for women and 3 years for men) the spurious positive employment effect in years 0-2 disappears.

In the fourth year after high school graduation, about 70 per cent of the individuals in the sample were employed (according to our definition of employment, which is employed for at least one month during the year and had positive earnings). From year four until year eight following high school graduation, the treatment effect estimates on employment are negative. The largest effect is about -0.02 in the natural experiment sample but it is not precisely estimated. When stacking the data for these four years (from the fourth to the seventh year following high school graduation), the estimate (-0.014 se=0.008) is statistically different from zero.[23] From year seven following high school graduation, the treatment effect on employment is positive, statistically significant in some years and marginally so in others. The highest employment treatment effect estimate is about 3 percentage points. The average employment rate from the seventh year following high school graduation is about 87 per cent and this rate is stable until the end of the period. The dynamic pattern obtained based on the RD sample is very similar.

The year-by-year estimated treatment effects on annual earnings are presented in Figures 4. These estimates are negative from the third to the sixth year following graduation, and then they turn positive and remain so until the end of the period studied. The lowest estimate based on the natural experiment sample is about -4,000 Israeli Shekels relative to mean earnings of 25,000 Israeli Shekels in the same year. The period with negative earnings effect coincides with the years with negative employment effect and with the period when the treatment effect on university enrollment becomes positive and increasing. This inverse of the treatment effect on employment and on university enrollment explains the negative effect on earnings. The treatment effect on earnings turns positive and significant from year seven on but it fluctuates in size, not surprisingly because earnings is a noisier outcome than university enrollment.

[23] However, it is likely that these estimates are attenuated because the employment measure we use does not account for part time employment and students usually have lower labor supply while in school.

Table 4, panel A, presents the DID estimates and their standard errors for the various labor market outcomes in the natural experiment sample. Column 2 presents the estimates at the end point of the period of study, eleven years after high school completion. The mean of each outcome based on the 2000 (pre-treatment) cohort is presented in column 1. I also present in columns 3-4 estimates from stacked regressions where I pool the data from the last three years of the period studied, namely nine to eleven years after graduating high school. The stacked regression yields the average treatment effect for this period and allows a more precise estimation of the effect on labor market outcomes. Focusing on the results from controlled regressions, based on outcomes measured 11 years after high school graduation (panel A), the teachers' incentive experiment increased treated students' employment rate by 1.0 percentage points and months of work by a third of a month. The relevant average employment rate is 84.1 per cent and the number of months of work is 9 months. The average unemployment rate in the treated group before treatment is low, only 6.8 per cent, which may be the reason why there is no discernible effect on this outcome. This rate is very similar to the national unemployment rate in 2010-2012 (7.1 per cent) for the closest age group (25-34). The estimated effect on eligibility for unemployment benefits is small and not significantly different from zero. The effect on annual earnings is positive, NIS 6,117, and it is statistically significant (se=2,963). This treatment effect on earnings amounts to a 9 per cent annual increase in earnings relative to the pre-program treatment group mean.[24]

The precision of the estimates of the effect on labor market outcomes declines when using the wild bootstrap procedure instead of the school level clustering. Both sets of standard errors for all labor market outcomes are presented in online appendix Table A8. Almost all estimates that are statistically significant based on school level clustering are also statistically significant based on the bootstrapped standard errors but at a lower level of significance. For example, based on the natural experiment sample estimates, the school level clustering yields a t-value for the effect on earnings 9-11 years after high school graduation of 2.22, while based on bootstrapped standard errors the t-value is 2.12. When using the RD sample, the t-value estimate obtained based on bootstrapped standard errors is 1.69 relative to 1.89 based on clustered standard errors. Therefore, the evidence on the effect of the program on earnings is less precise and therefore less conclusive in comparison to the estimated program effect on university schooling.

The stacked regression evidence (columns 3-4) is in line with the evidence presented in columns 1-2. The average earnings 9 to 11 years after high school completion is NIS 53,817 ($14,162). The average estimated effect of the PFP program on annual earnings for this three year period is NIS 4,714

[24] The estimates obtained when assigning the treatment status randomly to schools are all statistically not different from zero. These results are in online appendix Table A9. The estimates from the second robustness check where I estimated the DID model using the same samples of schools, but using earlier or later years as a non-treated period, are small and not significantly different from zero (online appendix Table A10).

(se=2,124), also 9 percent relative to the period mean earnings.[25] The estimates based on the RD sample are similar (columns 5-8). For example, the effect on annual earnings in the stack regression is NIS 4,860, se=2,576, also 9 percent of annual earnings (NIS 55,145). [26]

In panel B of Table 4 I present the *simple* differences in differences estimates for the labor market outcomes. These estimates are very close to those presented in panel A, reaffirming our earlier conclusion that the treatment and control groups are well balanced in characteristics and also in terms of outcomes of the pre-treatment cohort. For example, the simple difference in differences estimate on earnings 11 years after school completion is 5,552 (se=3,013) and the controlled difference in differences estimate is 6,117 (se=2,963). The respective estimated effects from the stacked regression for 9-11 years after school completion are 4,353 (se=2,107) and 4,714 (se=2,124).

In panel B of Table 3, I present labor market outcomes cross-section regressions separately for the 1999 and 2000 pre-treated cohorts, and for the 2002 post-treatment cohort, along with a similar regression for the 2001 (treatment) cohort, similar to those presented in Table 3 panel A for the post-secondary schooling outcomes. The 2001 post-treatment treatment-control difference in earnings is -1,026 and not statistically different from zero. On the other hand, the cross sectional differences in 1999, 2000, and 2002 (pre-treatment) cohorts are very similar and all three are different from the treatment-control differences in the 2001 (treatment) cohort. These differences are -5,893 for the 1999 (pre-treatment) cohort, -5,944 for the 2000 (pre-treatment) cohort, and -5,947 for 2002 (post-treatment) cohort. I note here again that these gaps are eliminated once we control for school type (Arab versus Jewish school) and therefore this comparison shows that the DID estimated effect on earnings in the 2001 (treatment) cohort is due to an improvement in outcomes of the treated group. Note also that the DID treatment estimate is the same regardless which year is used as pre-treatment.

A natural question about the above estimated effect on earnings is whether it captures the permanent long term effect. First, note that I measure the effect on earnings at about age 30-31 when individuals had already completed their post-secondary schooling. Second, based on a sample of older cohorts, I find that earnings at age 30-35 is a strong predictor of earnings at an older age. However, earnings have larger variation over time than other personal outcomes. To get a better indication of the permanency of the effect on earnings, I estimated the effect on the percentile rank of individuals in the distribution of their cohort. There is no direct evidence that suggests that rank forecast is more stable than earnings or log earnings. However, recent papers in the intergenerational mobility literature provide some indirect evidence that is relevant. These studies have shown that movements across ranks

[25] The fact that the treatment effect on earnings stays positive over several consecutive years is perhaps an indication that this gain reflects real productivity differences and not signaling of the higher schooling outcomes that resulted from 'teaching to the test'.

[26] The estimations based on the broader bandwidths instead of the 40-52 original bandwidth yield very similar results. For example, the estimated effect on earnings is 7,312 (se=3,367) in the 38-53 range and 7,204 (se=2,892) in the 37-54 range. These results are presented in online appendix Table A11. Note that these two estimates are only marginally different from the estimate obtained from the 40-52 per cent range, 7,093, se=3,677 (presented in column 6 of Table 4).

in the income distribution are uncorrelated with parental income conditional on rank at age 30; in contrast, movements in log earnings are correlated with parental income conditional on log income at age 30 - in particular, rich offspring have higher earnings growth, so that age 30 measurements are biased predictors of later-life earnings. However, the rank forecasts appear to be less biased. For example, Nybom and Stuhler (2016) show with data from Sweden that the relationship between a child's income rank and their parental income rank stabilizes by about age 30; in contrast, the relationship in log earnings is less stable. Chetty et al (2015) find a similar pattern in US tax data, reporting that percentile ranks predict well where children of different economic backgrounds will fall in the income distribution later in life. Using log earnings instead leads to inferior predictions because of the growth path expansions at the top of the income distribution.

Table 5 presents estimates of the effect of the program on percentile rank of earnings, where the rank is computed separately for each cohort. The estimates are fully consistent with the estimated effects on earnings that are presented in Table 4. Nine to 11 years after high school graduation, the program moves treated individuals in the natural experiment sample by 2 percentile ranks (column 4, first row) and this effect is relatively precisely measured (se=1.1). The rest of the estimates presented in the table suggest similar findings.[27]

5d. Treatment Heterogeneity by Family Earnings and Gender

Next, I estimate program treatment heterogeneity in university education, employment and earnings, by baseline family income. The possibility of a different program effect by family income has policy implications with respect to the targeted versus universal implementation of teachers' incentives programs, and for the external validity of our findings with respect to different socio-economic backgrounds of treated students. The sample is divided by the median of family income in 2000-2002. The estimates based on these two samples are presented in Table 6 for the natural experiment sample and p-values for group differences are presented as well in the table. For post-secondary schooling, I focus on outcomes 12 Years after High-School Graduation. For employment and earnings outcomes I focus on stacked regression estimates 9-11 years after high school graduation. Overall, the only significant differences in the effects of the program are between the high and low family income groups in the estimates of employment and months of work. All other differences are not precisely measured, likely due to the small size of the subsamples.

Panel A presents the estimates for sub-samples by family income. The effect on university ever enrollment and years of schooling is positive and significant for both groups but the effect in the high-income sample is twice as large as the effect in the lower income sample. Yet the standard errors of these estimates do not allow to conclude that they are statistically different from each other. The effect on earnings is larger for the low-income group, in contrast to the opposite gap in the gain in university

[27] The uncontrolled difference in differences estimates of the percentile rank regressions, not presented here for sake of space, are very similar to the control difference in differences estimates presented in Table 5.

schooling. However, what explains the larger gain in earnings in the low-income group is its larger increase in employment, both the employment rate and months of work. The estimated increase in the employment rate is 3.7 percentage points versus no employment effect at all for the higher family income sample. The estimated increase in months of work among this group is half a month (0.48, SE=0.181). However, it is interesting to note that the increase in earnings for a unit gain in university education is similar in the low and high family income samples.

The estimated effects by gender are presented in panel B of Table 6. The effect on schooling is positive and significant for both genders, but it is higher (though not statistically different given the estimated standard errors) for girls, a gain of a third of a year of university versus about half of that for boys. However, boys have a larger increase in earnings, first due to a larger positive effect on employment and secondly perhaps due to a higher rate of part-time work among women in this age group. Indeed, based on the 2012 Israeli Labor Force Survey, the rate of part time work among women in the age group 29-34 is 25 per cent versus 8 per cent among men.

The comparison between the treatment effect on men and women provides some evidence to support the claim that the differences in long-run earning between treatment and control really measure the effect of the treatment. Since men typically perform three years of military service in Israel while girls serve only two, we should expect the effect to start a year earlier for girls. The dynamic pattern of the effect on earnings presented in Figure 5 demonstrates that this is indeed the case. For both boys and girls, the effect in the first few years is zero because most of them are still in military service, but in the third year after high school graduation the earnings treatment effect for girls becomes negative (when boys are still serving in the army) and it is even more negative in year four. These negative effects are because of higher enrollment rate in university education, which reduces the likelihood of employment. The earnings treatment effect becomes positive only in year 7 when most girls finish university and commence employment. For boys, the drop to negative treatment effect occurs in the fourth year after high school completion and it becomes positive in year six after high school completion. The whole dynamic pattern for boys, which overall is similar to that of girls, displays a one-year lag, consistent with the additional year of military service for boys.

5e. Mechanisms for the Effect on Earnings

The direct effect of the program on high school outcomes, for example, the increase in the average composite score in the matriculation exams and the increase in the matriculation diploma rate, could have caused the increase in earnings that I find. Lavy, Ebenstein and Roth (2016) use random shocks to performance in matriculation exams to identify the reduced form effects of these high school outcomes on earnings at adulthood and find a strong and significant positive effect. While these results are similar in magnitude, it is important to independently analyze the sources of the reduced form effect of 6 to 9 percentage points increase in earnings that I estimated in this paper. First, I should account for the contribution of the positive effect on employment to the increase in earnings. In the natural

experiment sample, the employment gain accounts for 2 of the 9 percentage points increase. A second factor explaining earnings growth is the increase in university years of schooling. Recent estimates of the rate of return to a year of university education in Israel range from 12 to 16 per cent.[28] The lowest estimate (12%) implies that the 0.25 increase in years of university education contributed 3 percentage points to the gain in earnings. The highest estimate (16%) implies that the increase in university education accounts for 4 percentage points of the increase in earnings.

Another factor that accounts for part of the increase in earnings is the direct effect of the improved matriculation outcomes on earnings, independently of the effect they have on years of university education. Particularly important is the matriculation rate, which increased by 3.5-5 percentage points. For example, Angrist and Lavy (2009) estimate that *bagrut* holders earn 13 per cent more than other individuals with exactly 12 years of schooling. Therefore, the matriculation rate accounts for almost 0.5 percentage points of the earnings gain in the natural experiment sample. Similarly, the quality improvements in the matriculation study program (as reflected in the composite score, number of credit units and credits in honor and science subjects) are also rewarded in the labor market beyond their effect on post-secondary education(Caplan et al ,2009).[29] The implied mechanism is that the improvements in high school outcomes that resulted from the PFP intervention gave students access to higher quality post-secondary education, mainly by facilitating enrollment into more selective programs that have a higher return to education.

5f. Effect on Marriage and Children

I next examine teacher PFP treatment impacts on students' marriage and fertility outcomes in Table 7. I define these outcomes 11 years after graduating from high school because I only have data for these variables for 2012. Therefore, the two outcomes I examine are an indicator of being married and an indicator of having children, 11 years after high school graduation. About 58 per cent of the students from the pre-treatment treated schools sample are married by 2011. The treatment effect on marriage rate based on the natural experiment sample is negative, but not significantly different from zero. Similarly, the estimated effect on having children is negative but very small and not statistically different from zero. In panel B I report results based on sub-samples by family income. For the sub-sample of low family income, the estimates for marriage and children are negative and the first effect is large: a decline of 3.3 percentage points in the marriage rate, and a decline of 1.7 percentage points in the probability of having children, 11 years after high school graduation – but these effects are not precisely measured. The two estimates for the high family income sample are smaller and not

[28] Frish (2009) exploits changes in compulsory schooling laws and reports IV estimates much larger than OLS estimates. Navon (2005) estimates that the return to an MA degree (two years of schooling) is 30 percent.

[29] Caplan et al (2009) demonstrate that earnings in Israel are highly positively correlated with the quality of post-secondary education (colleges versus universities and higher versus lower quality universities). For example, it shows that earnings are much higher for graduates of the Tel Aviv, Hebrew and Technion Universities relative to graduates from the other four universities. Admission to the top universities is of course positively correlated with the high school matriculation outcomes.

statistically significant, but given their estimated standard errors we cannot conclude that they are different from the respective estimates of the lower income sample. In panel C, I report the results by gender, and clearly the treatment effect on the two demographic outcomes is small and not distinguishable from zero.

6. Conclusions

In this paper, I study the long term effect of an experiment that paid teachers a bonus based on their students' performance in high-stakes exams at the end of high school. All studies of teachers' incentive programs and the vast majority of published research on the impact of other school interventions has examined their effects on short-run outcomes, primarily by looking at their impact on standardized test scores. This study is the first to use a long horizon follow up, from high school to age 30, to examine the impact of a teachers' PFP scheme on long-term life outcomes. This analysis addresses the critical question of whether a public education intervention can achieve the ultimate goal of improving lifetime well-being. It also makes an important contribution to the growing literature on the long-term effects of education quality, by providing evidence about an intervention that changes a specific input which can improve student achievement. Focusing on an intervention that can be expanded or implemented elsewhere, such as teachers' PFP, provides explicit guidance for policymaking. This line of research is a natural follow up to recent studies that estimated a positive effect of teaching quality using teachers' fixed effect and value-added models, however, explicit policy prescriptions for how to improve teaching quality do not follow immediately from this important evidence. The results presented in this paper help in this regard by revealing an important element in the 'black box' of teacher quality.

This study shows that more than a decade after the initial intervention treated individuals experienced sizable gains in educational attainment and quality and large increases in annual earnings, some of which reflect a return to education quality beyond the return to years of schooling. Overall, the effect on post-secondary schooling is measured with more precision than the gain in labor market outcomes. However, these gains are very large relative to the cost of the program. The average cost of the program was $170 per student versus a gain close to $2,000 in annual earnings starting at about age 28-30. A complete cost-benefit analysis should also take into account the foregone earnings during post-secondary education and tuition fees. However, given that individuals will benefit for many years from the increase in earnings, the present value of benefits clearly outweighs the cost, suggesting a high private rate of return. A social rate of return analysis of this project should take into account the cost of university education not recovered by tuition fees, and the additional tax revenue levied on higher earnings. Clearly, these adjustments will still yield a high social internal rate of return on this project.

Merit and incentive based pay for teachers is being contemplated or implemented in many countries, making the evidence in this paper relevant and important for education policy worldwide. In U.S. education policy, for example, merit pay reforms for teachers returned to the top of the policy

agenda in the Obama Administration. In his first major education policy speech, President Obama promoted merit pay for teachers and in 2009 announced the Race to the Top, supported by $4.4 billion in federal funds, to encourage states to implement performance pay for teachers.[30] In a 2014 UK reform, teachers' annual salary increases have been tied to performance, replacing a system where almost all teachers automatically moved up a point on the pay scale every year. The move has been hugely controversial. For example, on March 26, 2014, the National Union of Teachers went on strike to protest the overhaul of pay structures that was due to begin later in the year.[31]

The intervention described in this paper targeted the period leading up to high-stake exams that play a key role in determining university and college admission. Since in this experiment the stakes were high both for students and teachers, it makes sense that the PFP intervention produced long-term results. However, if a similar program were introduced in a primary or middle school, the gains in test scores may not necessarily lead to similar long-term effects. Nevertheless, the evidence presented here is relevant for countries that use similar high-stakes, high school exams for university admission.[32] Another point to note is that a PFP program when implemented at scale, for example nationwide, will have general equilibrium implications. The scope of expansion of university enrollment estimated above will only be possible if the supply of post-secondary education can meet the increase in demand, as was the case in this study.

7. References

Abramitzky, R. and V. Lavy. 2014. "How Responsive is Investment in Schooling to Changes in Returns? Evidence from an Unusual Pay Reform in Israel's Kibbutzim", *Econometrica*, Vol. 82, No. 4 (July), 1241–1272.

Angrist, J. and V. Lavy, 1999, "Using Maimonides' Rule to Estimate the Effect of Class Size on Children's Academic Achievement." *Quarterly Journal of Economics*, (114) (May), 533-575.

Anderson, M. L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103 (484), 1481-1495.

Angrist, J. D. and A.B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106 (4), 979-1014.

Black, D. K. Daniel, and S. Sanders. 2002. "The Impact of Economic Conditions on Participation in Disability Programs: Evidence from the Coal Boom and Bust." *American Economic Review*, 92(1): 27-50.

[30] Merit-Pay For Teachers | eduflow: https://eduflow.wordpress.com/2013/10/08/merit-based-pay-for-teachers.
[31] The Economist, March 29 2014.
[32] High school exit exams play a similar role for university admission in many countires, for example, in the UK A level exams, the 'Bacalaureate' exams in Finland, Germany, Italy and Norway, or the 'Matura exams' in Austria, Switzerland, and most East European countries.

Caplan, Tom, Orly Furman, Dmitri Romanov, and Noam Zussman. 2009. “The Quality of Israeli Academic Institutions: What the Wages of Graduates Tell About It?” Central Bureau of Statistics, Israel, Working Paper Series NO. 42, May.

Card, David and Krueger, Alan B. 1991. “Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States.” *Journal of Political Economy, pp:*1-40.

Chetty, R., J. Friedman, N. Hilger, E. Saez, D. Whithmore Schanzenbach, and D. Yagan. 2011. “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *Quarterly Journal of Economics* 126(4): 1593-1660.

Chetty, R., J. N. Friedman, and J. Rockoff. 2014. “Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates” *American Economic Review*, 104(9): 2593-2632.

_____. 2014. “Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood” *American Economic Review*, September, 104(9): 2533-2679

Chetty, R., N. Hendren and L. Katz. 2016, “The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment”, *American Economic Review* 106(4): 855-902.

Chetty, R., N. Hendren, P. Kline, and E. Saez, 2014. “Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States” *Quarterly Journal of Economics* 129(4): 1553-1623.

Deming, David. 2009. “Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start," *American Economic Journal: Applied Economics*, 1 (3), 111-134.

Deming David, S. Cohodes, J. Jennings, and C. Jencks. 2016. “School Accountability, Postsecondary Attainment and Earnings.” *The Review of Economics and Statistics*, Dec. 98(5): 848–862.

Dustmann C., P. Puhani and U. Schonberg. 2012. “The Long-Term Effects of School Quality on Labor Market Outcomes and Educational Attainment”, CReAM Discussion Paper Series 1208.

Dynarski, S., J. Hyman, and D. Whitmore Schanzenbach. 2013. “Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion” *Journal of Policy Analysis and Management*, 32(4).

Duflo E., R. Hanna, and S. Ryan. 2012. “Incentives Work: Getting Teachers to Come to School” *American Economic Review*, pp: 1241-78.

Ebenstein A. Lavy V. and S. Roth. “The Long Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution”, *American Economic Journal: Applied Economics*, 2016, 8(4): 36–65.

Frish, R. 2009, “The Economic Returns to Schooling in Israel” *Israel Economic Review* (7) 113–141.

Fryer, R. G. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," *Journal of Labor Economics* Vol. 31, No. 2 (April), pp. 373-407.

Garces, E., D. Thomas, and J. Currie. 2002. “Longer-Term Effects of Head Start," *American Economic Review*, pp: 999-1012.

Glewwe, P., N. Ilias and M. Kremer. 2010. “Teacher Incentives*,” American Economic Journal: Applied Economics*, pp: 205-27.

Gould E., V. Lavy and D. Paserman. 2011. “Fifty-Five Years after the Magic Carpet Ride: The Long-Run Effect of the Early Childhood Environment on Social and Economic Outcomes”, *Review of Economic Studies*, July: 77, 1164–1191.

Jacob, B. A., and S. D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics*, 118, 843-77.

Johnson, R. C., C. K. Jackson and C. Persico, 2016. “The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms”, *The Quarterly Journal of Economics*, 131, 157–218.

Krueger, Alan B., and Diane M. Whitmore. 2001. “The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR,” *Economic Journal*, CXI: 1–28.

Lazear, E. 2000. “Performance Pay and Productivity,” *American Economic Review*, December.

Lazear, E. 2001. “Paying Teachers for Performance: Incentives and Selection,” Draft, August.

Lavy, V. 2002. “Evaluating the Effect of Teachers’ Group Performance Incentives on Students Achievements.” *Journal of Political Economy*, 10 (6), December: 1286–1318.

Lavy, V. 2007. “Using Performance-Based Pay to Improve the Quality of Teachers”, *The Future of Children*, 87-110.

Lavy, V. 2009. “Performance Pay and Teachers’ Effort, Productivity and Grading Ethics”, *American Economic Review*, Vol. 99, No. 5, December: 1979-2011.

Lleras-Muney, Adriana. 2005. “The Relationship between Education and Adult Mortality in the United States," *Review of Economic Studies*, 72, 189-221.

Ludwig, J. and D. L. Miller. 2007. “Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*, 122, 159-208.

Ludwig, J., G. J. Duncan, L. A. Gennetian, L. F. Katz, R. C. Kessler, J. R. Kling and L. Sanbonmatsu. 2013. “Long-Term Neighborhood Effects on Low-Income Families: Evidence from Moving to Opportunity”, NBER Working Paper 18772.

Navon, Guy, 2006. "Human Capital Heterogeneity: University Choice and Wages," MPRA Paper 9708, University Library of Munich, Germany.

Neal, D. 2011. "The Design of Performance Pay in Education," Handbook of the Economics of Education, Vol. 4, pp. 495-550.

Nybom M. and J. Stuhler. “Biases in Standard Measures of Intergenerational Income Dependence”. Draft, April 1, 2016.

Table 1: Desriptive Statistics and Tests For Balance Between Treatment and Contorl School Students in the Pre-Treatment Cohort (2000)

	Natural Experiment Sample			Regression Discontinuity Sample		
Dependent variable	Treated Schools (1)	Non-treated Schools (2)	Difference (3)	Treated Schools (4)	Non-treated Schools (5)	Difference (6)
A. Enrollment in Post-Secondary Schooling						
University	0.216	0.191	0.026 (0.046)	0.209	0.148	0.061 (0.039)
Academic College	0.145	0.143	0.001 (0.029)	0.160	0.127	0.033 (0.032)
B. Post-Secondary Years of Schooling						
University	0.825	0.716	0.109 (0.211)	0.793 (0.155)	0.523 (0.100)	0.270 (0.184)
Academic College	0.379	0.365	0.014 (0.080)	0.423	0.317	0.106 (0.084)
C. Employment Outcomes in 2011						
Employment (1 = Yes, 0 = No)	0.841	0.868	-0.027 (0.020)	0.842	0.868	-0.027 (0.021)
Months Worked	8.991	9.498	-0.507* (0.252)	9.077	9.559	-0.482* (0.236)
Annual Earnings (NIS)	62,001	67,682	-5,681 (3,265)	63,823	66,954	-3,131 (3,204)
Annual Unemployment Insurance Benefits (1 = Yes, 0 = No)	0.068	0.073	-0.005 (0.009)	0.069	0.073	-0.004 (0.010)
Annual Unemployment Insurance Benefits (NIS)	693	754	-61 (107)	725	711	15 (115)
D. Demographic Outcomes						
Married (1 = Yes, 0 = No)	0.563	0.542	0.020 (0.042)	0.548	0.558	-0.010 (0.050)
Children (1 = Yes, 0 = No)	0.451	0.454	-0.003 (0.051)	0.427	0.487	-0.060 (0.053)
Number of Children	0.836	0.792	0.044 (0.127)	0.760	0.830	-0.070 (0.138)
Age at First Marriage	24.338	24.591	-0.253 (0.379)	24.586	24.495	0.090 (0.409)
Age at First Birth	25.302	25.467	-0.165 (0.357)	25.523	25.405	0.117 (0.415)

E. Parental Outcomes

Father Earnings in 2000-2002	102,212	96,212	6,001 (16,693)	103,816	81,924	21,892 (16,668)
Mother Earnings in 2000-2002	47,715	45,484	2,231 (7,798)	49,082	39,383	9,699 (6,945)
Number of Observations	2,424	2,703	5,127	1,697	2,471	4,168
Weighted Number of Observations	3,980	4,171	8,151	2,843	3,064	5,907

Notes: The table reports means and treatnebt-control differences for different Post-Secondary schooling, employment, income, and dempographic variables. Columns 1-3 report results based on the natural experiment sample, and columns 4-6 report results based on the regression discontinuity sample. Panel A presents the binary variables indicating whether the individual has been enrolled in a university or college, by 2012. The categories are not mutually exclusive and overlapping is possible. Panel B reports the number of years of education an individual has attained by 2012 by university and college. Panel C reports means of employment and income outcomes in 2011. Panel D reports mean of demographic variables in 2012. Panel E reports means of parental earnings in 2000-2002. Standard errors in parenthesis are adjusted for school level clustering. Number of observations does not apply to the the age at marriage and the age at first birth variables becasue these are computed for individuals that married/have children by 2012. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 2: Estimates of the Effect of Teachers' Bonuses Program on Post-Secondary Schooling (12 Years After High-School Graduation)

	Natural Experiment Sample				Regression Discontinuity Sample			
	Enrollment in Post-Secondary Schooling		Post-Secondary Years of Schooling		Enrollment in Post-Secondary Schooling		Post-Secondary Years of Schooling	
	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Controlled Difference in Differences								
University	0.216	0.048**	0.825	0.250**	0.209	0.060***	0.793	0.242**
	(0.412)	(0.019)	(1.877)	(0.094)	(0.407)	(0.020)	(1.856)	(0.103)
Academic College	0.145	-0.026	0.379	-0.072	0.160	-0.017	0.423	-0.047
	(0.352)	(0.019)	(1.051)	(0.052)	(0.367)	(0.026)	(1.107)	(0.066)
Any Post-Secondary Schooling	0.522	0.028	1.922	0.170*	0.517	0.041*	1.952	0.191*
	(0.500)	(0.020)	(2.366)	(0.089)	(0.500)	(0.022)	(2.390)	(0.092)
Number of Observations	2,703	10,077	2,703	10,077	2,471	8,230	2,471	8,230
Weighted Number of Observations	4,171	15,903	4,171	15,903	3,064	11,561	3,064	11,561
B. Uncontrolled (Simple) Difference in Differences								
University	0.216	0.039*	0.825	0.206**	0.209	0.047*	0.793	0.182
	(0.412)	(0.021)	(1.877)	(0.102)	(0.407)	(0.023)	(1.856)	(0.107)
Academic College	0.145	-0.028	0.379	-0.075	0.160	-0.021	0.423	-0.055
	(0.352)	(0.020)	(1.051)	(0.053)	(0.367)	(0.026)	(1.107)	(0.064)
Any Post-Secondary Schooling	0.522	0.019	1.922	0.123	0.517	0.023	1.952	0.111
	(0.500)	(0.022)	(2.366)	(0.097)	(0.500)	(0.025)	(2.390)	(0.103)
Number of Observations	2,703	10,077	2,703	10,077	2,471	8,230	2,471	8,230
Weighted Number of Observations	4,171	15,903	4,171	15,903	3,064	11,561	3,064	11,561

Notes: This table presents the differences-in-differences estimates of the effect of the teachers' bonus program on Post-Secondary schooling 12 years after high-school graduation. Panel A reports results for the controlled DID and Panel B for the uncontrolled simple differences. Columns 1-4 report the results based on the natural experiment sample and columns 5-8 based on the regression discontinuity sample. Columns 1,3,5, and 7 present the means and standard errors for the 2000 (untreated) cohort in the treated schools and it is used as benchmark for assessing the size of the treatment effect. Columns 2,4,6, and 8 report the Differences-in-Differences estimates for each of the dependent variables. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Cross Section Estimates, Outcomes 12 Years After High-School Graduation, Specification With Controls

	Natural Experiment Sample				Regression Discontinuity Sample			
	1999	2000	2001	2002	1999	2000	2001	2002
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Post Secondary Education								
Ever Enrolled University	-0.035	0.017	0.055	0.000	0.001	0.036	0.105	0.058
	(0.029)	(0.032)	(0.039)	(0.034)	(0.026)	(0.034)	(0.038)	(0.032)
Ever Enrolled College	-0.032	0.004	-0.019	-0.034	-0.008	0.028	0.013	-0.002
	(0.025)	(0.028)	(0.025)	(0.025)	(0.034)	(0.032)	(0.033)	(0.030)
Total Years of Schooling University	-0.084	0.058	0.254	0.040	0.061	0.137	0.409	0.260
	(0.130)	(0.146)	(0.171)	(0.137)	(0.128)	(0.156)	(0.170)	(0.128)
Total Years of Schooling College	-0.060	0.024	-0.031	-0.046	0.007	0.089	0.066	0.067
	(0.076)	(0.075)	(0.078)	(0.076)	(0.098)	(0.084)	(0.096)	(0.086)
B. Employment and Income								
Employment Indicator (1 = Yes, 0 = No)	-0.020	-0.021	-0.010	-0.040	-0.034	-0.020	-0.008	-0.033
	(0.013)	(0.019)	(0.014)	(0.019)	(0.014)	(0.020)	(0.019)	(0.023)
Months Worked	-0.218	-0.405	-0.095	-0.492	-0.366	-0.430	-0.020	-0.513
	(0.169)	(0.228)	(0.203)	(0.224)	(0.189)	(0.224)	(0.246)	(0.270)
Total Annual Earnings (NIS)	-5,893	-5,944	-1,026	-5,947	-3,293	-4,649	255	-5,374
	(2,532)	(2,325)	(2,390)	(2,441)	(2,618)	(2,362)	(2,657)	(2,409)
Percentile Ranking of Total Annual Earnings (NIS)	-2.186	-2.944	-0.460	-1.774	-1.475	-2.537	0.250	-1.706
	(1.230)	(1.528)	(1.346)	(1.116)	(1.203)	(1.569)	(1.454)	(1.247)

Notes: This table presents the cross-sectional difference in Post-Secondary schooling outcomes 12 years after high school graduation. Panel A presents resimates for for post secondary education outcomes, in columns 1-4 based on the natural experiment sample and in columns 5-8 based on the regression discontinuity sample. Panel B presents estimates for employment and income outcomes, also for both samples. Standard errors are clustered at the school level.

Table 4: Estimates of the Effect of The Teachers' Bonuses Program on Employment and Income

	The Natural Experiment Sample				The Regression Discontinuity Sample			
	11 Years After High-School Graduation		9-11 Years After High-School Graduation, Stacked Regression		11 Years After High-School Graduation		9-11 Years After High-School Graduation, Stacked Regression	
	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Controlled Difference in Differences								
Employment (1 = Yes, 0 = No)	0.841	0.010	0.830	0.013	0.842	0.008	0.832	0.012
	(0.366)	(0.013)	(0.376)	(0.012)	(0.365)	(0.010)	(0.373)	(0.009)
Months Worked	8.988	0.321*	8.670	0.234	9.078	0.393**	8.763	0.194
	(4.605)	(0.172)	(4.670)	(0.142)	(4.549)	(0.164)	(4.620)	(0.137)
Annual Earnings (NIS)	64,993	6,117**	53,817	4,714**	66,903	7,093*	55,145	4,860*
	(56,317)	(2,963)	(48,411)	(2,124)	(57,493)	(3,677)	(49,072)	(2,576)
Annual Unemployment Insurance Benefits (1 = Yes, 0 = No)	0.068	0.000	0.070	-0.002	0.069	-0.004	0.071	-0.005
	(0.252)	(0.015)	(0.255)	(0.006)	(0.254)	(0.019)	(0.257)	(0.008)
Annual Unemployment Insurance Benefits (NIS)	693	37	597	28	725	-112	602	-14
	(3,076)	(160)	(2,699)	(60)	(3,193)	(186)	(2,694)	(82)
Number of Observations	2,703	10,077	8,109	30,231	2,471	8,230	7,413	24,690
Weighted Number of Observations	4,171	15,903	12,525	47,745	3,064	11,561	9,195	34,695
B. Uncontrolled (Simple) Difference in Differences								
Employment (1 = Yes, 0 = No)	0.841	0.008	0.829	0.012	0.842	0.007	0.832	0.012
	(0.366)	(0.013)	(0.376)	(0.012)	(0.365)	(0.009)	(0.373)	(0.009)
Months Worked	8.988	0.297*	8.670	0.216	9.078	0.369**	8.763	0.178
	(4.605)	(0.164)	(4.670)	(0.135)	(4.549)	(0.157)	(4.620)	(0.130)
Annual Earnings (NIS)	64,993	5552*	53,817	4353**	66,903	6352*	55,145	4411*
	(56,317)	(3,013)	(48,411)	(2,107)	(57,493)	(3,590)	(49,072)	(2,480)
Annual Unemployment Insurance Benefits (1 = Yes, 0 = No)	0.068	0.001	0.070	-0.001	0.069	-0.001	0.071	-0.004
	(0.252)	(0.015)	(0.255)	(0.008)	(0.254)	(0.020)	(0.257)	(0.010)
Annual Unemployment Insurance Benefits (NIS)	693	45	597	34	725	-87	602	-4
	(3,076)	(160)	(2,699)	(83)	(3,193)	(189)	(2,694)	(110)
Number of Observations	2,703	10,077	8,109	30,231	2,471	8,230	7,413	24,690
Weighted Number of Observations	4,171	15,903	12,525	47,745	3,064	11,561	9,195	34,695

Notes: This table presents the differences-in-differences estimates of the effect of the teachers' bonuses program on employment and income outcomes. Panel A reports results for the controlled DID and Panel B for the uncontrolled simple differences. Columns 1-4 report the results based on the natural experiment sample, and columns 5-8 based on the regression discontinuity sample. Columns 1-2 and 5-6 report results for 11 years after high-school graduation, and columns 3-4 and 7-8 report results based on regressions with stacked data of 9-11 years after high-school graduation. The 'Employment' outcome equals 1 if an individual has worked at least one month during the year and had positive earnings, 0 otherwise. The outcome 'Annual Unemployment Insurance Benefits' equal 1 if the individual recieved any positive amount of unemployment benefits in the given year, 0 otherwise. The outcome 'Annual Unemployment Insurance Benefits' equals the NIS amount of unemployment benefits an individual received in a given year. Columns 1,3, 5, and 7 report the mean and standard error for the 2000 cohort (untreated) in the treated schools and it is used as benchmark for assessing the size of the treatment effect. Columns 2,4, 6, and 8 report the Differences-in-Differences estimate for each of the dependent variables listed above. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Estimates of the Effect of The Teachers' Bonuses Program on Percentile Ranking of Income

	The Natural Experiment Sample				The Regression Discontinuity Sample			
	11 Years After High-School Graduation		9-11 Years After High-School Graduation, Stacked Regression		11 Years After High-School Graduation		9-11 Years After High-School Graduation, Stacked Regression	
	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Controlled Difference in Differences								
Annual Earnings (NIS)	48.777	2.638**	48.3	2.041*	49.814	3.018**	49.0	1.972
	(30.337)	(1.210)	(30.3)	(1.114)	(30.547)	(1.423)	(30.5)	(1.256)
Number of Observations	2,703	10,077	8,109	30,231	2,471	8,230	7,413	24,690
Weighted Number of Observations	4,171	15,903	12,513	47,709	3,064	11,561	9,192	34,683

Notes: This table presents the differences-in-differences estimates of the effect of the teachers' bonuses program on income percentile ranking. Percentile ranking of income is assigned within each cohort and are age-adjusted. Columns 1-4 report the results based on the natural experiment sample, and columns 5-8 based on the regression discontinuity sample. Columns 1-2 and 5-6 estimates are based on data for 11 years after high-school graduation, and columns 3-4 and 7-8 estimates are based stacked data for outcomes 9-11 years after high-school graduation. Columns 1,3, 5, and 7 report the mean and standard error for the 2000 cohort (untreated) in the treated schools and it is used as benchmark for assessing the size of the treatment effect. Columns 2,4, 6, and 8 report the Differences-in-Differences estimate for each of the dependent variables listed above. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Estimates by Family Income and by Gender, 11 Years After High-School Graduation for Post-Secondary Schooling and 9-11 Years After High-School Graduation for Stacked Employment and Income Regressions - The Natural Experiment Sample

	University Enrollment		University Years of Education		Employment		Months Worked		Annual Earnings	
	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Controlled Difference in Differences										
A. By Family Income										
High Family Income	0.269	0.067*	1.053	0.344**	0.861	-0.009	9.492	0.012	58,562	3,521
	(0.444)	(0.036)	(2.043)	(0.154)	(0.346)	(0.014)	(4.246)	(0.157)	(50,832)	(3,133)
Low Family Income	0.132	0.029	0.450	0.130*	0.789	0.037**	8.364	0.480**	47,876	6,149**
	(0.339)	(0.021)	(1.362)	(0.065)	(0.408)	(0.017)	(4.886)	(0.181)	(44,495)	(2,316)
P-value		0.4125		0.2192		0.0297		0.053		0.4802
B. By Gender										
Boys	0.187	0.028	0.703	0.161*	0.848	0.025	9.497	0.257	60,574	5,646
	(0.390)	(0.021)	(1.710)	(0.084)	(0.359)	(0.020)	(4.328)	(0.233)	(51,954)	(3,694)
Girls	0.230	0.066	0.870	0.333*	0.810	0.001	8.465	0.196*	46,778	2,976
	(0.421)	(0.039)	(1.881)	(0.167)	(0.392)	(0.016)	(4.763)	(0.184)	(43,319)	(2,326)
P-value		0.4578		0.3855		0.4029		0.8427		0.5528

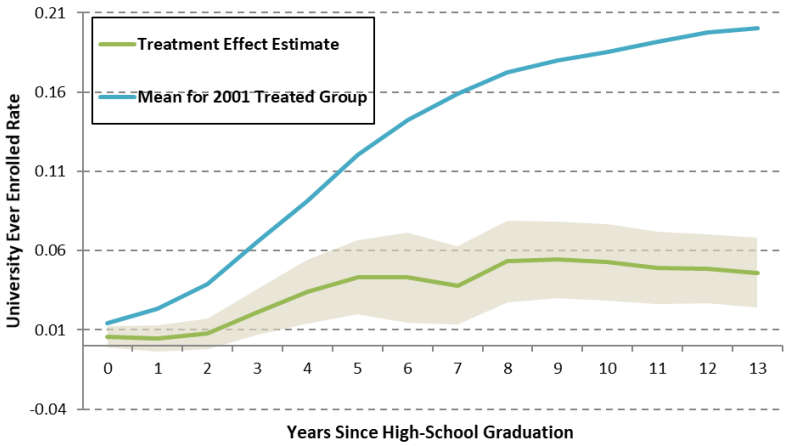
Notes: This table presents the differences-in-differences estimates of the effect of the teachers' bonuses program on Post-Secondary schooling, employment, and income, by family income and by gender and the p-values of differnces between the estimates in each subgroup. Results are based on the natural experiment sample. Panel A reports the results for the high and low family income, respectively. High family income is defined above the mean household income in 2000-2002. Panel B reports the results for boys and for girls, respectively. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Estimates, Demographic Outcomes 11 Years After High-School Graduation - The Natural Experiment Sample

	Married		Children	
	Mean of 2000 Cohort in Treated Schools	Estimate	Mean of 2000 Cohort in Treated Schools	Estimate
	(1)	(2)	(3)	(4)
Controlled Difference in Differences				
A. Full Sample				
	0.584	-0.011	0.451	-0.003
	(0.493)	(0.018)	(0.498)	(0.015)
B. By Family Income				
High Family Income	0.554	-0.003	0.407	0.006
	(0.497)	(0.027)	(0.491)	(0.020)
Low Family Income	0.621	-0.033	0.506	-0.017
	(0.485)	(0.024)	(0.500)	(0.025)
P-value	-	0.17	-	0.50
C. By Gender				
Boys	0.493	-0.005	0.338	-0.003
	(0.500)	(0.029)	(0.473)	(0.024)
Girls	0.677	-0.007	0.569	-0.001
	(0.468)	(0.022)	(0.495)	(0.023)
P-value	-	0.76	-	0.954

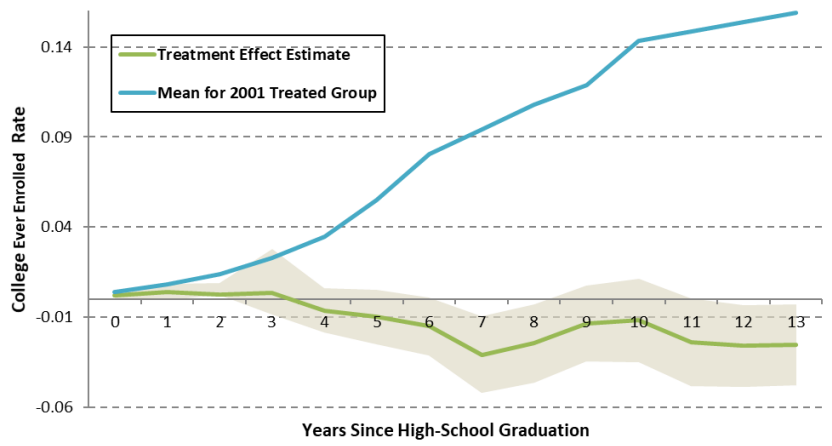
Notes: This table presents the DiD estimates of the effect of the teachers' bonuses program on different demographic outcomes 11 years after high-school graduation based on the natural experiment sample. Panel A reports the results for the full sample, Panel B reports the results for the high and low family income samples, and Panel C reports the results by gender. Columns 1-2 report the results for the 'Married' outcome which equals 1 if an individual is married 11 years after graduation, 0 otherwise. Columns 3-4 report the results for the outcome 'Children' which equal 1 is an individual has any children by 11 years after graduation, 0 otherwise. Columns 1 and 3 report the mean and standard error for the the 2000 cohort (untreated) in the treated schools and it is used as benchmark for assessing the size of the treatment effect. Columns 2 and 4 report the differences-in-differences estimate for each of the dependent variables listed above. Standard errors are clustered at the school level.

Figure1: Mean and Treatment Effect: University Enrollment (Full Sample, Natural Experiment Sample)



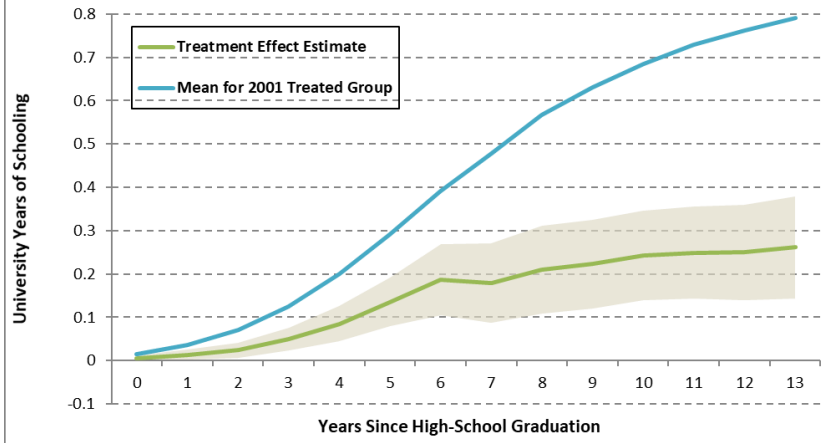
*Shaded area indicates two sided confidence intervals, 10% significance level.

Figure 2: Mean and Treatment Effect: College Enrollment (Full Sample, Natural Experiment Sample)



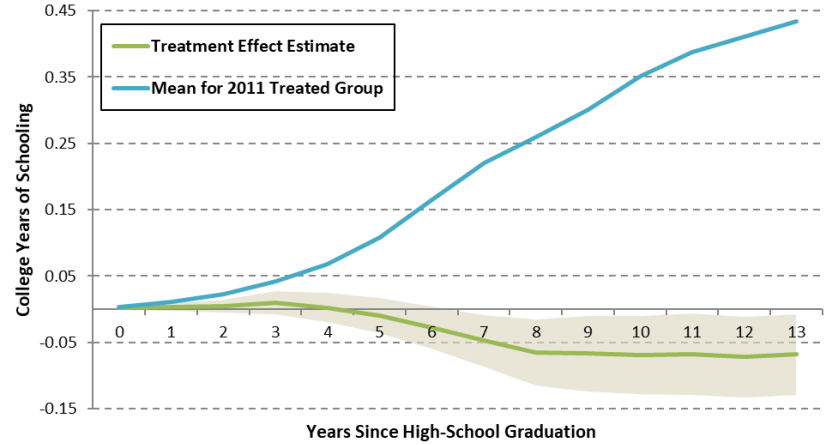
*Shaded area indicates two sided used confidence intervals, 10% significance level.

Figure 1A: Mean and Treatment Effect: University Years of Schooling (Full Sample, Natural Experiment Sample)



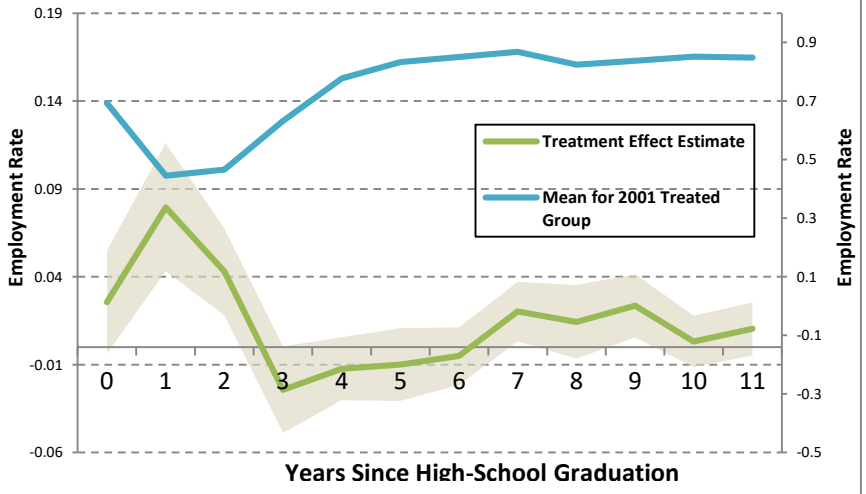
*Shaded area indicates two sided used confidence intervals, 10% significance level.

Figure 2A: Mean and Treatment Effect: College Years of Schooling (Full Sample, Natural Experiment Sample)



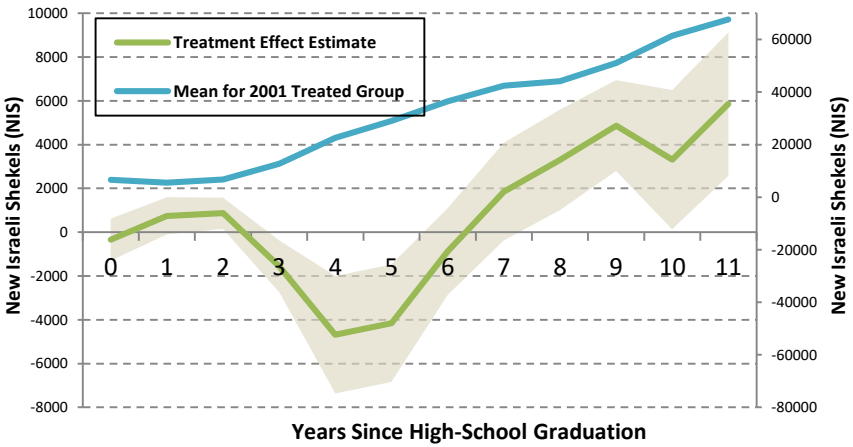
*Shaded area indicates two sided used confidence intervals, 10% significance level.

Figure 3: Mean and Treatment Effect: Employment Rate (Full Sample, Natural Experiment Sample)



*Shaded area indicates two sided confidence intervals,

Figure 4: Mean and Treatment Effect: Annual Earnings - 2009 Prices NI Shekels (Full Sample, Natural Experiment Sample)



*Shaded area indicates two sided confidence intervals,

Figure 5: Treatment Effects on Annual Earnings (2009 Prices), By Gender