NBER WORKING PAPER SERIES

THE EFFECT OF HIGH SCHOOL
MATRICULATION AWARDS: EVIDENCE FROM
RANDOMIZED TRIALS

Joshua D. Angrist
Victor Lavy

The Effect of High School Matriculation Awards:
Evidence from Randomized Trials
Joshua D. Angrist and Victor Lavy
NBER Working Paper No. 9389
December 2002
JEL No. I21, I28, J13, J24

<u>ABSTRACT</u>

In Israel, as in many other countries, a high school matriculation certificate is required by universities and some jobs. In spite of the certificate's value, Israeli society is marked by vast differences in matriculation rates by region and socioeconomic status. We attempted to increase the likelihood of matriculation among low-achieving students by offering substantial cash incentives in two demonstration programs. As a theoretical matter, cash incentives may be helpful if low-achieving students reduce investment in schooling because of high discount rates, part-time work, or face peer pressure not to study. A small pilot program selected individual students within schools for treatment, with treatment status determined by previous test scores and a partially randomized cutoff for low socioeconomic status. In a larger follow-up program, entire schools were randomly selected for treatment and the program operated with the cooperation of principals and teachers. The results suggest the Achievement Awards program that randomized treatment at the school level raised matriculation rates, while the student-based program did not.

Joshua D. Angrist
MIT Department of Economics
50 Memorial Drive
Cambridge, MA 02142-1347
and NBER
angrist@mit.edu

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

One of the most economically important education milestones in many countries and in some American states is a high-school matriculation exam. Examples include the French Baccalaureate, the New York State Regents examinations, and the recently instituted Massachusetts Comprehensive Assessment System. In Israel, a national high school matriculation exam -- known as the Bagrut – is a pre-requisite for admission to universities and arguably marks the dividing line between the working class and the middle class. In spite of the Bagrut's economic and social value, Israeli society is marked by vast differences in Bagrut completion rates across regions and by socioeconomic status. These disparities have led Israeli educators and administrators to try remedial programs in an attempt to increase high school matriculation rates. But most of these programs appear to have had little effect. Similarly, an array of service-oriented anti-dropout demonstrations for American teens appears to have had little effect on graduation rates (Dynarski and Gleason, 1998).

The discouraging results from previous anti-dropout interventions stimulated our interest in a simpler approach that focuses on immediate financial incentives for student effort. As a theoretical matter, cash incentives may be helpful if low-achieving students have high discount rates, reduce investment in schooling by going to work, or face peer pressure not to study. The promise of more immediate financial rewards may tip the scales in favor of schoolwork. In this paper, we report on two demonstration projects that provided financial awards for low-achieving high school students in Israel. The interventions discussed here rewarded Bagrut completion and performance on Bagrut subject tests with direct payments to students. We also discuss some of the methodological issues arising in evaluations of this type.

The task of evaluating educational incentives presents a number of practical and research-design challenges. How should the incentives be structured? Are incentives that target entire schools more likely to be effective than those that target individual students? What is an appropriate experimental design? To shed light on these questions, we explored two alternatives. The first was a small pilot study that randomized students within schools. Because simple random assignment within schools was seen as inequitable by administrators, we used a hybrid within-school experimental design. Eligibility for treatment was first

determined by forecasting the probability of Bagrut completion as a function of socioeconomic characteristics and previous test scores. Only students with performance in a certain range were offered treatment. Random assignment was introduced by lowering the selection threshold with probability one-half, a modification justified to school administrators on the grounds that the program budget had insufficient funds to treat everyone below the original threshold.

A second larger intervention involved 40 schools with a very low percentage of high school seniors obtaining Bagrut certification. This program differed from the pilot program primarily in that treatment assignment was determined by random assignment of entire schools and not students within schools. The school-level awards program was also implemented with the cooperation of principals and administrators in treated schools. Another difference is that intermediate awards were offered for completion of individual subject tests and for continued school enrollment, though the highest awards were offered to seniors who ultimately obtained a Bagrut.

Random assignment of schools rather than students generates a group-randomized trial (GRT) of the type widely used to study interventions in naturally clustered units such as schools, hospitals, and communities (see, e.g., Donner, Brown, and Brasher, 1990). GRTs offer practical and cost-saving advantages, but may not balance treatment and control characteristics when, as is typical, a small number of units are randomized. Under some circumstances, balance can be improved by using matched pairs, as in our experimental design. Another important disadvantage of clustered designs is that, because outcomes within clusters are correlated, GRTs usually have much lower statistical power than simple randomized trials with the same sample size (see, e.g., Feng, Diehr, Peterson, and McLerran, 2001). Conclusions may also depend on the statistical framework used to interpret GRT data; in particular, whether to treat the group or the individual as the basis for inference. We therefore explore a number of approaches to the analysis of data from the schools GRT.

Section I provides background on the Israeli school system and describes the school-based

achievement awards program in detail. This section also outlines the theoretical context for programs of this type. Other interventions in this spirit include the Quantum Opportunities Program (QOP; Hahn, Leavitt, and Aaron, 1994); the Learning, Earning, and Parenting (LEAP) demonstration project in Ohio (Long, Gueron, Wood, Fisher, and Fellerath, 1996); the Education Maintenance Allowance (EMA) in Britain (Deardon, *et al*, 2001); Progresa in Mexico (Behrman, Sengupta, and Todd, 2000; Schultz, 2001); and the Programa de Ampliación de Cobertura de la Educación Secundaria (PACES), which provided private school vouchers in Colombia (Angrist, Bettinger, Bloom, King, and Kremer, 2002).[1] The Achievement Awards experiment also has elements in common with the college tuition subsidy programs run by the I Have a Dream Foundation, and Robert Reich's (1998) proposal to pay targeted bonuses of $25,000 to high school graduates from low-income families.

Section II discusses the pilot experiment and Section III presents background for the school-based demonstration. Section IV discusses results from the schools demonstration, which, while not entirely clear-cut, suggest the probability of Bagrut certification increased by 6-8 percentage points in Award schools. Consistent with a causal interpretation of these results, analyses conditional on previous test results show treatment effects only for students with achievement levels that put them in a position to benefit from additional effort. In contrast, randomization within schools clearly failed to generate a change in achievement. Together these results support the hypothesis that the participation of school administrators and teachers is important for the success of individual incentive programs like those tested here. Section V concludes and outlines directions for further work.

---

[1]QOP combined services for children in AFDC families with modest financial incentives for enrollment in a small randomized pilot. LEAP used financial incentives along with case management and support services to increase the school enrollment of welfare mothers in a randomized demonstration. EMA pays children or mothers of children in low-income families based on enrollment and achievement, and is currently being evaluated in a non-randomized study. Progresa offered payments based on the enrollment status of primary and secondary school children in randomly selected towns in Mexico. The PACES program awarded vouchers for private school in a lottery for 6th graders in Colombia. This had an achievement component because vouchers were lost if students failed to keep up with schoolwork. As far as we know, however, ours are the first demonstrations to combine substantial achievement-based payments to students with a randomized experimental design.

I. Background

A. Theoretical Context

Why do young men and women fail to complete high school? Why don't more people go to college? These questions present something of a puzzle since the economic returns to schooling appear to be very large, and almost certainly exceed the costs of additional schooling for most non-college graduates. Research on education choices suggests possible explanations for low schooling levels, mostly related to heterogeneity in costs (or perceived costs) and heterogeneity in returns (or expected returns). Using data from the NLSY, for example, Eckstein and Wolpin (1999) link the drop-out decision to lack of ability and motivation, low expectations about the rewards from graduation, disutility from schooling, and a comparative advantage in the jobs available to non-graduates. Another consideration raised in the literature on college attendance is liquidity constraints and the role of financial aid (see, e.g., Fuller, Manski, and Wise, 1982; Card and Lemieux, 2000). Since capital markets are imperfect and human capital is hard to collateralize, some poor students may chose not to go school in the absence of subsidies.

A number of features of the Israeli economic environment dovetail with the issues raised in previous research on low educational attainment. First, while high school is free, there is an opportunity cost to schooling since students can work, perhaps at the expense of remedial programs that might make Bagrut success more likely. A related concern is that some teenagers act as if they have very high discount rates (see, e.g., Gruber, 2000). Israeli requirements for compulsory military service (at least 3 years for boys and 2 years for girls) probably exacerbate the impact of discounting since working life for a male college graduate does not begin until 6-7 years after high school. Uncertainty about returns may also be greater for poor Israelis, who are disproportionately likely to live in small towns with few educated adult role models. Finally, peer effects may be a negative influence in some of the relatively isolated communities where education is lowest.

Bagrut status in Israel is not directly comparable to an American student's drop-out status since most

4

of the students who fail to complete a Bagrut still finish their secondary schooling. Nevertheless, as for American high school dropouts, post-secondary schooling options for Israeli graduates without a Bagrut are limited; very few non-Bagrut holders will obtain further schooling. Of course, many students may not be able to complete a Bagrut no matter how hard they try. But the substantial cross-sectional and time series variation in Bagrut rates suggests that some students attending schools with low completion rates could, under some circumstances, do better. This possibility is highlighted by the Ministry of Education's practice of reporting the proportion of high school seniors who are "close" but fail to obtain a Bagrut, on the order of 22 percent.[2]

The Achievement Awards demonstrations were motivated by a desire to tip the scales towards current investment in schooling and away from market work or leisure, especially for students close to the margin for Bagrut success. The concrete and relatively immediate awards offered by the programs should have increased the present value of studying for exams and reduced uncertainty about returns. The programs may also have provided a cover story that students could use to justify schoolwork in the face of ridicule by non-studying peers. Our intervention is in the spirit of Reich's (1998) proposal to offer students from low-income families in the US a $25,000 cash bonus for graduating high school. Keane and Wolpin (2000) simulated the impact of this policy in the context of a structural model of education choice. They estimated that the Reich program would have a large impact on high school graduation rates and college attendance, especially for Blacks.

B. The Israeli School System

Israeli education consists of elementary school (grades 1-6), middle school (grades 7-9) and high school (grades 10-12). High school students are enrolled in an academic track leading to a Bagrut, or in a

---

[2]Bagrut statistics for the 2000 school year are available at http//www.netvision.net.il/bagrut/netunim2000.htm#1.4.

vocational track leading to a diploma. The Bagrut is completed by passing a series of national exams in core and elective subjects beginning in 10th grade, with more tests taken in 11th grade and most taken in 12th grade. Students choose to be tested at various proficiency levels, with each test awarding 1 to 5 credit units per subject, depending on difficulty. Some subjects are mandatory and many must be taken for at least 3 units.[3] A minimum of 20 credit units is required to qualify for a matriculation certificate. About 52 percent of all high-school seniors received a matriculation certificate in the 1999 and 2000 cohorts (Israel Ministry of Education, 2001). Roughly 60 percent of those who took at least one Bagrut subject test end up receiving a Bagrut certificate. In our samples, however, Bagrut rates are much lower.

II. The Pilot Experiment

Our investigation of financial incentives for students began with a pilot evaluation in the 1999-2000 school year involving approximately 500 students in low-achieving schools in Southern Israel. We hoped to increase the number of students taking the various Bagrut subject tests, to encourage students in trouble to take advantage of remedial services, and to increase study effort at school and at home. Students in the treatment group were told that Bagrut achievers would have their choice of $800 in cash, a $1,000 voucher to be used towards a trip of educational value, or a $1,200 voucher to be used towards the cost of higher education. Students who were offered the opportunity to earn an award were notified in writing in March 2000. The time lines for both the pilot and follow-up programs appear in the Appendix.

Treated students in the pilot were notified of the intervention later than originally planned, a fact that reduced the scope for changes in behavior that could have increased the likelihood of obtaining a Bagrut in June. On the other hand, students have the opportunity to boost their efforts towards the Bagrut at any time,

---

[3]Bagrut subject requirements change from year to year and are described in the appendix. Some Bagrut tests are graded internally, but internal grades that deviate substantially from external scores are disqualified. The Achievement Awards program, which neither awarded nor sanctioned teachers, would seem to have offered little incentive to lower standards and risk disqualification.

and have access to remedial instruction through schools and outside of school. Moreover, Israeli seniors have the opportunity to try some Bagrut subject tests again the following winter, provided they have not already been drafted. The pilot data analyzed here include the winter retests.

The pilot experiment differed from the larger school-based experiment which followed in that we used a design that randomized students within schools. Randomization within schools should have provided a more powerful design with less data. In practice, however, simple randomization was seen by school administrators as hard to justify to participants and outside observers. We therefore used a hybrid instrumental variables design that selected students for treatment on the basis of their socioeconomic characteristics and a partially randomized threshold, described in more detail below.[4]

A. Sample Selection, Experimental Design and Descriptive Statistics

The sample was selected by dividing the population of 1302 seniors enrolled in the 1999-2000 school year into three groups on the basis of the number of Bagrut subject tests they had taken previously and their maximum score on these tests (recall that the Bagrut is determined by a series of tests, some of which are usually taken in 10th and 11th grade). We estimated Logit regressions with information from the previous cohort of students to predict the probability of Bagrut certification as a function of these two variables, denoted here by p_{1i} for student i. All students with a very low probability of Bagrut attainment ($p_{1i}<.053$) were offered the opportunity to earn a bonus. It was inexpensive and politically expedient to offer bonuses to this group, about 15 percent of enrolled seniors in the Southern cohort.

At the other end, students with a very high probability of success were excluded, in particular, we did not offer bonuses to 612 students with $p_{1i}>.66$, about half of seniors. The remaining 491 students were

_____

[4]Azgursky and Schmidt (2001) compare an instrumental variables design and clustered experimental design in a simulation study.

7

potentially eligible for an award.[5] Treatment was assigned to these students as a function of family size and father's education, with students of lower socioeconomic status more likely to be in the treatment group. We chose this mechanism because the Israeli Ministry of Education commonly allocates education resources in this way (see, e.g., Angrist and Lavy, 1999). Again, we used the previous cohort of seniors to estimate the probability a student would obtain a Bagrut certificate as a function of family size and father's schooling, denoted p_{2i}. To introduce an element of random assignment, we used a threshold that varied by student and school, so that roughly half of the eligible students from each school were offered treatment. The rule for treatment assignment for student i in school j was

$$T_{ij} = 1[p_{2i} < q_{.22}(j)(1-Z_i) + q_{.7}(j)Z_i]$$

where $q_{.22}(j)$ and $q_{.7}(j)$ are the .22 and .7 quantiles of the p_{2i} distribution in school j.

The hybrid experimental designed is sketched in Table 1, which reports the number of students in the all-treated, no-treated, and eligible samples. For the eligible sample, the table shows the number of students exposed to a high (Z_i=1) and low (Z_i=0) threshold and the number offered the opportunity to earn an award. The design is such that about half of eligible students were exposed to a high threshold and about half of eligibles had the opportunity to earn an award. Exposure to the high threshold instead of the low increased the probability of treatment by about 50 percentage points since the probability of treatment increased from about one-quarter to about three-quarters.

Descriptive statistics for the eligible, all-treated, and non-treated samples are presented in Table 2. About one-third of eligible students received a Bagrut.   By construction, the probability of receiving a Bagrut is very low (%2.7) in the all-treated sample, and relatively high (%77.3) in the no-treated samples. Other covariates in the table are the test-history variables used to construct the index determining eligibility status (i.e., the regressors used to construct p_{1i}) and the socioeconomic regressors used to construct p_{2i}. The

---

[5]The analysis sample includes 489 eligible students because of missing demographic data not used for treatment assignment but used in the analysis.

bottom panel provides information on some additional characteristics that can be used to assess the success

of the random assignment of the threshold shifter, Z_i. There is no significant association between Z_i and any

covariate. There is also no significant association between Z_i and Bagrut status, although those exposed to

a high threshold were much more likely to have been given the opportunity to earn an award.

The experimental design detailed in Tables 1 and 2 amounts to simple random assignment for those

with p_{2i} in the interval $[q_{.22}(j), q_{.7}(j)]$. Alternately, we can think of this as random assignment conditional on

the covariates p_{2i} and school effects. Finally, the design can be seen as generating an instrument, Z_i, for the

endogenous regressor, T_{ij} (endogenous in this case means correlated with family background). We focus on

the instrumental variables interpretation, reporting reduced form estimates of the effects of exposure to the

high threshold on the likelihood of obtaining a Bagrut certificate.

B. Results

Table 3 reports the reduced form estimates for models including a range of covariates. Data are from

the eligible sample of 489 and a subsample of 439 that excludes students attending the single Bedouin school

in the sample. Some of the models include school fixed effects. Although exposure to a high threshold is

associated with a precisely estimated 50-53 percentage point first stage, there is no association between Z_i

and Bagrut attainment. Note that because the first-stage effect was .5, standard errors for the effect of

treatment are approximately double those for the reduced form effects shown in the table.[6]

In exploratory analyses, we found some evidence of a positive effect on Bagrut rates for girls, though

also puzzling negative effects for boys. Both results are significantly different from zero. The positive effect

for girls seems plausible, but the negative effect for boys is puzzling, especially since it indicates a

remarkably large decline in Bagrut rates for boys in the treatment group. A possible explanation is that this

---

[6]The hybrid design was powerful enough to detect a treatment effect of about .12. Because the control group Bagrut rate was higher in the pilot sample, this is roughly the same proportional treatment effect found in the clustered design.

particular analysis by subgroup, one of many possible analyses of this type, uncovered this pattern by chance. As a specification check, we therefore repeated the analysis by sex in two further subgroups. The first is the "random assignment sample", i.e., those students with p_{2i} in the interval $[q_{.22}(j), q_{.7}(j)]$. For this group $T_{ij}=Z_i$, and the effect of Z_i should be strongest. In contrast, for those with p_{2i} outside this interval, the "no-first-stage sample", treatment status is orthogonal to Z_i and determined solely by whether p_{2i} is below $q_{.22}(j)$ or above $q_{.7}(j)$.

The results of this specification check, reported in Table 4, suggest the negative effect for boys and positive effect for girls is just a chance occurrence. Estimates using the no-first-stage sample show no relationship between Z_i and treatment status for either sex, yet the association between Z_i and Bagrut rates is even larger than in the random assignment sample where $Z_i=T_{ij}$. We therefore conclude that the pilot experiment had no effect on achievement, though a null hypothesis of modest effects cannot be rejected either. This built-in specification check is a useful feature of the hybrid design. Another useful finding is that of the 80 awards given out in the pilot, only 2 were to students from the low-scoring all-treated sample. No award recipients chose the travel option, and only 5 chose the $1,200 tuition voucher. We therefore switched to an all-cash scheme in the larger school-based experiment.

III. Background for the Schools Experiment

To kick off the school-based demonstration, we conducted an orientation with principals and administrators from treatment schools in January 2001. Some principals chose not to participate, though most were enthusiastic, and informed their students shortly thereafter, usually in a school assembly or a classroom announcement. Many schools also distributed written materials describing the program to students and/or their parents.[7] The award schedule is detailed in the appendix. The program was meant to last 3

---

[7]We interviewed principals at each treatment school to verify this. Unlike in the pilot, treated students in this case were notified early enough to request a deferment for military service if they wanted to retake Bagrut tests. In practice, this option seems unlikely to have been a major contributor towards program effects.

years, with awards given to high school students in every grade. Two awards were offered to students who progressed from 10th to 11th grade and from 11th to 12th grade. Small awards of NIS500 were also given for test-taking regardless of the outcome, with NIS1500 given for actually passing tests before senior year. The largest award was NIS6,000 (almost $1,500) for any senior who received a Bagrut.

The total amount at stake for a student who passed all achievement milestones was NIS10,000 or just under $2,400. This is about one-third of the after-tax earnings a student could expect from working full-time as a high-school drop-out, and about twice as much as a student might earn working full-time in two summer months. Due to adverse publicity, however, the awards program was suspended after the first year. The suspension was announced in May of 2001, about a month before the Bagrut tests. As a consequence, awards were given for only one year of achievement and the maximum amount awarded was NIS6,000. The suspension and associated public controversy should have reduced the program impact least for seniors, since they would have been in the program for only one year anyway.[8]

Given this unanticipated deviation from the intended scenario, it seems worth asking how likely the program is to have affected student behavior. As part of the follow-up effort, the Ministry of Education's evaluation division surveyed students in October 2001 to determine whether they remembered the program and whether behavior changed as a result. The response rate among seniors was low since many had already been drafted or were hard to locate for other reasons. Low-achieving students are probably over-represented so the survey results are suggestive at best. Nevertheless, almost 53 percent of the students interviewed recalled specific program features, and over 80 percent of these recalled attending a school assembly where program information was distributed. Among those who remembered the program, 87 percent said the bonus was large enough to induce extra effort and about half reported they did indeed work harder. We also

---

[8]In May 2000, when 2000 Bagrut results were announced, Education Ministry officials referred reporters to the Achievement Awards program as an attempt to increase scores. This led to extensive and mostly critical media coverage. The program was then suspended, though the Ministry issued a press release indicating the program would run as planned for the first year and then be assessed.

found that students in treated schools reported studying 2.7 hours per week between January and June, 11 percent more than the 2.4 average hours of study in control schools. This is consistent with the program having caused 25% of students to study an *extra* 2 hours per week for 3 months. The academic value of this extra effort is a separate question, however, and the subject of our impact evaluation.[9]

A. Experimental Design and Descriptive Statistics

In December 2000, we selected 40 high schools with low 1999 Bagrut rates, but above a minimum threshold rate of 3 percent. Some schools with low completion rates were ineligible to participate in the experiment for technical or administrative reasons. The list of participating schools included 10 Arab and 10 Jewish religious schools.[10] Treatment was randomly assigned to 20 of the 40 participating schools. The total number of treated schools was determined by the program budget constraint, which allowed about 750,000 dollars for award payments.

Random assignment of entire schools does not balance treatment and control characteristics as effectively as random assignment of students within schools. Nevertheless, while not large enough to ensure treatment-control balance, the number of clusters used here is typical (see, e.g., Feng, *et al*, 2001). To improve treatment-control balance we used a matching strategy that paired treatment and control schools based on lagged values of the primary outcome of interest, the average 1999 Bagrut rate.[11] Treatment was assigned randomly within pairs, as is common in GRTs (see, e.g., Gail, *et al*, 1996). Such pre-treatment matching is typically worthwhile provided that (a) matching effectively balances pre-treatment outcomes; and (b) lagged outcomes are a reasonably good predictor of future outcomes.

---

[9]In June of 2001, around the time of the Bagrut tests, an Israeli television station ran a special program that included interviews with pilot participants, as well as with one of us (Lavy) and program critics. The participants' comments suggested the program was of considerable interest to students.

[10]Israel runs semi-autonomous school systems for Secular Jews, Religious Jews, and Arabs. Rules for Bagrut are similar in all three systems.

[11]We used 1999 Bagrut rates to select and match schools because the 2000 data were incomplete when treatment was assigned.

Table 5 presents descriptive statistics for each of the 39 schools that were initially involved in the experiment. There are 39 schools instead of 40 because the control school in pair 6 had closed by the time treatment were assigned. In follow-up contacts in March 2001, we verified the level of program participation by contacting principals and school administrators. School administrators in two non-compliant schools (in pairs 14 and 15) hoped to participate but submitted student rosters shortly after the deadline. The principals of three schools had taken no concrete actions to inform students or teachers about the program and/or indicated that they did not wish to participate.

The enrollment figures in Table 5 show the number of high school seniors in each school year from 1999-2001. Religious schools tend to be smaller than secular schools. Schools range in size from 10 to 242 seniors in 1999, and some schools show marked changes in size from year to year. These changes reflect the unstable environment that characterizes Israel's weakest schools. Many absorb large cohorts of new immigrants and are in small towns with substantial population movements.

Bagrut rates in 1999 ranged from 3.6-28.6 percent. As noted above, this is much lower than the national average of 52 percent for high school seniors. It is important to note, however, that Bagrut rates in 2000 and 2001 were much more variable than those in 1999. This partly reflects our sample design since rates for 1999 were selected to be within a certain range. The variability in Bagrut rates in later years also results from small school size, changes in school populations due to immigration and internal migration, and measurement error in the Bagrut data. In practice, the 1999 Bagrut rate is not as powerful a predictor of the 2000 and 2001 Bagrut rates as we had hoped. The R^2 from a weighted regression of the 2001 rate on the 1999 rate is .15. On the other hand, the overall Bagrut rate in the sample was reasonably stable, ranging from 20-22 percent.

Although the variability documented in Table 1 is clearly undesirable, it bears emphasizing that substantial variability in year-to-year performance measures for individual schools is not unique to our sample. Using data on North Carolina schools, for example, Kane and Staiger (2001) similarly found that

13

much of the year-to-year variation in performance is due to school-level random shocks that come from sources other than sampling variance and permanent differences in school characteristics.

Table 5 also reports the probability of being on the Bagrut track for seniors at each school. Most of the students in the sample were registered as being on the Bagrut track in spite of the low probability of ultimately receiving a Bagrut. In the analysis that follows, we focus on samples that include all students since reported track-status may be endogenous. This endogeneity is a consequence of the fact that track status is reported with error, and errors are more likely to be corrected for those students who ultimately received a Bagrut.

B. Econometric Framework

Because treatment was randomly assigned in the schools experiment, unbiased estimates of treatment effects can be obtained from simple treatment-control comparisons. In practice, however, a number of complications are worth special attention. First, as noted above, randomization by cluster is less likely than individual-level randomization to balance confounding factors, even after matching. This is especially relevant in view of the unstable Bagrut rates in Table 5. We attempted to improve treatment-control balance by discarding 4 pairs with the largest (as measured by t-statistics) differences in 2000 Bagrut rates. Other econometric issues are discussed below.

Adjusting for Non-Compliance

Schools' compliance status may be endogenous in the sense that it was partly determined by anticipated Bagrut rates. If so, estimates in a sample limited to schools that complied will be biased. A simple approach to the compliance problem is to estimate "intention-to-treat effects", i.e., the reduced-form impact of the randomized offer of program participation in the full sample. Such estimates are reported below. Intention-to-treat effects provide a lower bound on the effect of actual program participation and can

14

be re-scaled into effects on students in treated schools by using the randomized opportunity to participate in the program as an instrumental variable for actual participation. Because no control schools received treatment, this approach estimates the effect of treatment on the treated (Imbens and Angrist, 1994).

A second adjustment for compliance is suggested by Table 5. Note that if we could identify compliant schools *ex ante*, i.e. before treatment was assigned, efficient estimation procedures would limit the analysis to treatment/control pairs where the treatment school is compliant. Restricting the sample to compliant schools generates efficiency gains because this restriction exploits prior knowledge about the link between assignment and treatment. Compliance status is only known *ex post*, however, and is therefore endogenous. On the other hand, Table 5 shows that non-compliant schools are concentrated at the upper end of the distribution of 1999 Bagrut rates. Limiting the analysis to schools with 1999 rates less than .25 eliminates 3 out of 5 non-compliant schools and 2/3 of non-compliant students. We therefore report results for a "low-rate sample" of schools with 1999 Bagrut rates less than .25, as well as for the full sample and the balanced sample noted above. As it turns out, treatment and control schools are also more comparable (as measured by 2000 Bagrut rates) in the low-rate sample.

Inference in Group Randomized Trials

Randomized trials that assign treatment status to entire schools may be more attractive than within-school randomization for both programmatic and logistical reasons. First, school randomization reduces the perception of unfairness associated with randomization. Second, students not offered treatment may nevertheless be affected by treatments received by others in the same school, diluting within-school treatment effects. Finally, education interventions may be more effective when introduced at the school level. Incentive programs for students depend partly on the cooperation of teachers and school administrators, and may get additional leverage from peer effects when those nearby participate.[12]

---

[12]A recent experiment in financial education illustrates this point (Duflo and Saez, 2002).

15

The most important statistical issue in school-level GRTs is whether to treat groups (schools) or students as the unit of observation, and, if the latter, how best to adjust inferential procedures for clustering at the group level. As Cornfeld (1978) notes, analyses of GRTs that ignore clustering are "an exercise in self-deception." The traditional cluster adjustment relies on a linear model with random effects, an approach known to economists primarily through the work of Moulton (1986). When the clusters are all of size n, this amounts to multiplying standard errors by a "design effect," $[1+(n-1)\rho]^{1/2}$, where ρ measures the intra-cluster residual correlation. A problem with random effects models in this context is that the equi-correlated error structure they impose is implausible for binary outcomes like Bagrut status. Another problem is that estimates of ρ tend to be too low.

A modern variation on random effects models is the Generalized Estimating Equation (GEE) framework developed by Liang and Zeger (1986). GEE allows for an unrestricted correlation structure and can be used for binary outcomes and nonlinear models such as Logit and Poisson regression. An advantage of GEE is that it is very flexible and increasingly available in proprietary software.[13] The primary disadvantage is that the validity of GEE inference turns on an asymptotic argument based on the number of clusters (as do parametric random effects models). GRTs often have too few clusters for asymptotic formulas to provide an acceptably accurate approximation to the finite-sample sampling distribution.[14]

A simple alternative to micro-level analyses is to work with grouped data, in this case, school means. Average Bagrut rates are approximately Normally distributed so t-tests may be valid even with a moderate number of groups. On the other hand, grouped analyses are conservative in the sense that they treat additional observations within schools as if they were uninformative beyond their impact on the dispersion of the averages. Statistical tests based on grouped data may therefore be less powerful than those based on micro data.

---

[13]GEE standard errors are produced by the Stata "Cluster" option and SAS GENMOD procedure.
[14]See, e.g., Thornquist and Anderson (1992).

We present regression estimates of treatment effects using both micro and grouped data. The model used to construct estimates using student data can be written:

$$y_{ijt} = \alpha_j + x_{jt}{'}\beta + \sum_q d_{qi}\mu_q + \delta Z_{jt} + \epsilon_{ijt}, \tag{1}$$

where i indexes students; j=1, . . ., 20 indexes pairs; t=0,1 indexes treatment status within pairs, and Z_{jt} is *assigned* treatment status. Grouped covariates includes pair effects, α_j, and two school-level covariates denoted x_{jt}, which consist of dummies for Arab and religious schools. The student-level covariates are three dummies (d_{qi}; q=2, 3, 4) indicating the quartile of a student's average test score on Bagrut and diploma tests taken as of January 2001, when the program was implemented. It turns out that this lagged score variable, described in more detail below, is the single best predictor we have of students' Bagrut status.

Grouped estimates were constructed similarly. The typical grouped equation can be written:

$$\bar{y}_{jt} = \alpha_j + x_{jt}{'}\beta + \delta Z_{jt} + \bar{\eta}_{jt}, \tag{2}$$

where $\bar{y}_{jt}$ is the school average Bagrut rate and $\bar{\eta}_{jt}$ is a grouped error term. Some of these estimates are weighted by the school size, n_{jt}. As suggested by classical results on regression efficiency, it seems natural to weight, and weighted estimation using groups produces the same estimates as micro data when there are no covariates. On the other hand, in models with group random effects (implicit in this case since we worry about clustering), weighted estimation need not be more efficient. Moreover, when treatment effects are heterogeneous, weighted and unweighted procedures estimate different average effects.

We also experimented with a two-step procedure discussed by Baker and Fortin (2001) and Donald and Lang (2001). In our case, this amounts to adjusting school means for micro covariates by estimating school fixed effects in an equation like (1), and then regressing the estimated fixed effects on treatment status and other school-level covariates in an equation like (2). In particular, we first estimate

$$y_{ijt} = \mu_{jt} + \sum_q d_{qi}\mu_q + \epsilon_{ijt}, \tag{3}$$

and then regress $\hat{\mu}_{jt}$, the estimated μ_{jt}, on the same covariates as in equation (2). This procedure uses the micro-data to reduce the dispersion in means, while inference is conservative in the sense that no credit is

17

taken for within-cluster variability in the second step.

Donald and Lang (2001) present Monte Carlo evidence suggesting the two-step estimator has good finite sample properties for some designs and always improves on cluster-adjustments, though Baker and Fortin (2001) report second-step estimates that are sensitive to weighting. To address this point, we report weighted and unweighted second step estimates. Finally, in a direct attack on the problem of downward-biased GEE standard errors, we estimated standard errors using Bell and McCaffrey's (2002) Biased Reduced Linearization (BRL) estimator for micro data. BRL implements a correction for GEE standard errors similar to MacKinnon and White's (1985) bias-corrected heteroscedasticity-consistent covariance matrix. Bell and McCaffrey present Monte Carlo evidence suggesting BRL generates tests of the correct size in traditional random effects models.

A final statistical issue worth noting is that some of the regression estimates are from models that omit pair effects. In principle, pair effects can be dropped without biasing the estimates of treatment effects since intention to treat is assigned with equal probability across pairs. Ignoring stratification variables may also lead to more precise estimates in paired experiments since the inclusion of pair effects uses up degrees of freedom (Diehr, *et al*, 1995; Angrist and Hahn, 1999). On the other hand, with few pairs, a chance association between pair characteristics and treatment status is possible.

IV. Results from the Schools Experiment

Post-treatment Bagrut rates in treated schools are higher on average than those in control schools, conditional on baseline (2000) Bagrut rates. This can be seen in Figure 1, which plots 2001 Bagrut rates against 2000 Bagrut rates, with solid dots representing treated schools, and separate regression lines drawn through treatment and control observations. The plot incorporates Bagrut data from all 39 schools involved in the experiment and shows residuals from regressions on Arab and religious school dummies. Although the figure indicates that average Bagrut rates in 2001 were somewhat unevenly dispersed, the regression line

18

running through the averages for treated schools is almost everywhere above the regression line running through the averages for control schools.

Figure 2 plots the relationship between 2001 and 2000 Bagrut rates by treatment status, after regression-adjusting for pair effects as well as dummies for Arab and religious schools. Here the dispersion in 2001 rates is more uniform. The regression lines are necessarily parallel for this specification since the residuals for each pair sum to zero. But Figure 2 suggests that conditional on 2000 Bagrut rates, treated schools were likely to have higher 2001 Bagrut rates than control schools. The difference between the two lines in Figure 2 is about 8.5 percentage points.

A. Estimates Using School Means

As suggested by the figures, unweighted contrasts in school means show higher 2001 Bagrut rates in treated than control schools, with no corresponding difference in 2000. This can be seen in the first three columns of Table 6, which report estimates of equation (2) with no controls, adding school covariates (dummies for religious and Arab schools), and including school covariates and pair effects. For example, the uncontrolled difference in 2001 Bagrut rates is .075 in the full sample, though the standard error is almost as large at .063. Adding controls for school and pair effects increases the difference to .082, with a standard error of .059. At the same time, treatment-control differences in 2000 are all negative, a specification check that reinforces the causal interpretation of the 2001 results.[15]

The standard errors quoted above (and reported in Table 6 directly below the coefficient estimates) are conventional least-squares estimates, while those in brackets are heteroscedasticity-corrected. The fact that the corrected standard errors are substantially below the unadjusted standard errors, with the gap increasing in the number of covariates, suggests downward bias in the corrected estimates (see, e.g., Chesher and Jewitt, 1987). We therefore take the unadjusted standard errors as a more reliable measure of precision

---

[15]Results for 1999 Bagrut rates are not shown since these are balanced by the experimental design.

for the grouped estimates.

The balanced and low-rate subsamples generate larger treatment effects than the full sample, again with no evidence of a treatment-control difference in 2000 data. On the other hand, results from weighted contrasts in means, reported in columns 4-6, are less clear cut. With no controls, the weighted estimates are the same in the 2001 and 2000 full sample, a finding that clearly raises questions about the 2001 results. The picture is somewhat clearer, however, with additional controls and in the balanced and low-rate samples. For example, the weighted estimate with school covariates and pair effects in the balanced sample is .061 (s.e.=.043) in 2001 and .021 (s.e.=.03) in 2000. The weighted estimate for the balanced sample with school covariates only, reported in column 5, is .089 (s.e.=.047) in 2001 and .053 in 2000 (.045).

Note that the generally larger estimated effects (weighted or unweighted) in the balanced and low-rate samples are consistent with the fact that the compliance rate is 75 percent in the full sample, but 86-87 percent in the balanced and low-rate samples. Thus, we would expect intention-to-treat effects to be about 15 percent larger in the latter two samples, a factor that does not seem too far off the mark. Finally, note that while the estimates for 2001 increase as we move to the balanced and low-rate samples, this is not typically the case using 2000 data, providing an encouraging specification check.

B. Estimates Using Student Data

In an attempt to check robustness and increase the precision of the estimated treatment effects, we used micro data to control for students' performance on tests taken as of the baseline date, January 2001. In particular, we divided the credit-unit-weighted average score on all Bagrut and Diploma tests (coding zeros for those with no tests), and then coded dummies for each quartile of the score distribution. We used quartile dummies instead of, say, linear control for lagged scores, to facilitate the analysis conditional on lagged scores discussed below. The quartile dummies are a powerful predictor of students' ultimate Bagrut status. For example, the probability of being awarded a Bagrut 2001 was about 1% in the lowest quartile,

20

9% in the second quartile, 29% in the third quartile, and 49% in the upper quartile.

Adjusting for baseline scores using the two-step procedure described above generates more precise and mostly larger treatment effects than the analysis of group means. For example, the unweighted estimate from a model with pair effects and school covariates, reported in column 3 of Table 7, is .12 (s.e.=.051). The weighted estimate falls to .068 (s.e.=.041), but this now contrasts with an estimate for 2000 of only .039 (s.e.=.047). Moreover, the weighted estimates from this specification in the balanced and low-rate samples show significant treatment effects for 2001 on the order of .07-.08, with no corresponding effect in 2000. The weighted estimates in the balanced sample also point to a treatment effect when estimated in models with no controls and school covariates only.[16]

The last two columns in Table 7 report estimates of treatment effects in micro data. The standard errors reported in parentheses use a conventional GEE cluster-adjustment, while BRL standard errors are shown in brackets. The estimates in column 8 are all significant, even when precision is measured using the larger BRL standard errors. Estimates in column 7 for the balanced sample are also significant, and again on the order of 8%. Moreover, none of the micro-data estimates show evidence of a (spurious) treatment effect in 2000. Interestingly, the BRL standard errors are often close to the unadjusted two-step standard errors, typically slightly lower, though occasionally slightly higher.

On balance, the results in Tables 6 and 7 support the notion that the Achievement Awards program increased Bagrut rates in 2001 by something on the order of 6-8 percentage points (using the smaller weighted or micro-data estimates). As an additional check on the causal interpretation of these results, we estimated models allowing treatment effects to vary with lagged score quartiles. That is, we estimated

$$y_{ijt} = \alpha_j + x_{jt}{}'\beta + \sum_q d_{qi}\mu_q + \sum_q \delta_q Z_{jt} + \epsilon_{ijt}, \tag{4}$$

---

[16]The decline in standard errors when going from grouped to two-step estimates is consistent with fact that the standard deviation of the estimated $\hat{\mu}_{jt}$ is about 82 percent of the standard deviation of $\bar{y}_{jt}$. Note that the ratio of two-step to grouped standard errors for the weighted full sample is .041/.05 =.82 for the model in column 6. When constructing standard errors for the two-step estimator we ignore the fact that the micro coefficients, μ_q, are estimated using the full sample and therefore the estimated fixed effects are correlated. Since about 1000 students are available to estimate each quartile effect, this seems likely to be of minor importance.

where δ_q is a quartile-specific treatment effect and μ_q is a quartile main effect. Students with very low scores were unlikely to be able to obtain a Bagrut no matter how hard they tried in the treatment year. On the other hand, some relatively high-scoring students had scores in a range where extra effort may have made a difference. We therefore look for significant estimates of δ_3 and δ_4 in 2001, but not in 2000.

` As noted earlier, about half of students in the upper quartile ended up obtaining a Bagrut while almost no one in the lower quartile did. This can be seen in the pattern of control group means by quartile, which are reported along with quartile-specific treatment effects in Table 8. The model used for all of the estimates in this table included school covariates and pair effects, corresponding to the estimates in column 8 of Table 7. Small and insignificant treatment effects were estimated in the first two quartiles, with much larger and statistically significant estimates in the third and fourth quartiles. Results for 2000 show no significant effects for any quartile, though the coefficient estimates for 2000 are mostly larger in the third and fourth quartiles than in the first and second. The absence of a significant effect for any quartile in 2000 and the large positive and significant effects for the upper quartiles in 2001 support the view that the Achievement Awards program increased Bagrut rates.

The data used for Tables 6 and 7 and the first 4 columns of Table 8 come from the June 2001 round of Bagrut tests, i.e., before the Winter retests. We focused on the initial round of test results in the schools experiment because, as discussed above, the program was disrupted in early summer by adverse publicity. This seems likely to have reduced the scope for a treatment effect in the retests. A second consideration is that there was a unexpected round of second-chance Bagrut tests offered in math and English in late summer/early Fall 2001, between the first round and the traditional Winter round. We are not sure how this might have affected the Achievement Awards program, but some observers noted that the purpose here seemed to be to get Bagrut rates up by easing standards. In any case, the last 4 columns of Table 8, which report estimated treatment effects by quartile using data that incorporates results from the unexpected second chance and the Winter retests, show results broadly similar to the June results. The largest treatment control

differences appear in the third and fourth quartile, while there are no significant effects in 2000.[17] Estimates for the fourth quartile are somewhat smaller and no longer significant in the full sample, but remain significant in the balanced and low-rate samples. Estimates for the third quartile are somewhat larger.

V. Conclusions and Further Work

Although the evidence is not seamless, the school-based randomized trial suggests worthwhile gains in matriculation rates can be obtained by offering cash awards in low-achieving schools. The value of the awards was substantial from the point of view of high school seniors, but pales in comparison with the likely economic benefits of a matriculation certificate. To see this, note that the bonus offer of NIS6,000 shekels was worth about $1429 at the time the treated cohort finished school. About 27 percent of the treatment group received bonuses, so the cost was about $385 per treated student. To provide a rough assessments of the benefits, note that earnings of workers with 11-12 years of schooling in 2000 were about $16,100 (Israel central Bureau of Statistics, 2002). Those with some college earned 53 percent more. Suppose the causal effect of a Bagrut is less than half of this, say 25 percent, and that the effect of the program was to raise Bagrut rates by 7 percentage points. Then the program should increase annual earnings in the treated group by $16,100 \times .25 \times .07 = \$282$ per person *per year*, so the cost of the bonus will be quickly recovered.

Another way to benchmark costs and benefits is by comparison with other Bagrut-enhancement strategies. In the introduction we noted that most service-oriented strategies appear to have been ineffective. Not long after the Achievement Awards demonstration, however, the Ministry of Education piloted a relatively expensive service-oriented strategy offering intensive after-school instruction to small groups of under-performing students in several matriculation subjects. The results of an evaluation point to an 11 percent increase in Bagrut rates for students in the group offered treatment, at an average cost of $1,100 per student (Israel Ministry of Education, 2002). The after-school program therefore cost almost 3 times as

---

[17]The estimates for 2000 in columns 5-8 differ from those in columns 1-4 because, for comparability, the 2000 results used to construct these estimates also include Winter re-tests.

much, while producing an effect only about 50 percent larger.

In contrast with the school-based bonuses intervention, a smaller pilot program that selected individual students within schools for treatment, with treatment status determined by previous test scores and a randomized cutoff for low socioeconomic status, had no effect. This suggests that school-wide mobilization may be an important part of the incentive structure in programs of this type. On the other hand, delays and a relatively small sample may also account for the lack of a finding in the pilot.

On the methodological side, the paper compares alternative strategies for inference in a GRT. A graphical analysis suggests the program had an effect, albeit heterogeneous and variable across schools. Statistical analyses of school means generates results with a clear pattern of effects when unweighted, but more mixed results when weighted. A two-step method that uses micro data to reduce the dispersion of group averages generates somewhat sharper weighted results than a straight grouped analysis, while bias-corrected standard errors for micro-data estimates also leave an impression of significant effects. Finally, conditioning on lagged test scores generates a pattern of estimates consistent with the notion that the program raised matriculation rates.

The analysis here covers the immediate short-run impact on the Achievement Awards programs' target objective, high school matriculation status. In work in progress, we are looking at additional mediating outcomes, such as whether treated students took more tests or changed subjects, and whether they participated in remedial instruction. In future work, we hope to assess the long-run effects of the Achievement Awards program by collecting information on university attendance and possibly earnings. Finally, we are conducting a detailed Monte Carlo study of alternate modes of inference with school-based randomized trials.

REFERENCES

Angrist, Joshua, Eric Bettinger, Erik Bloom, Beth King, and Michael Kremer, "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review* 92 (2002), 1535-1558.

Angrist, Joshua, and Jinyong Hahn, "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," NBER Technical Working Paper No. 241, May 1999.

Angrist, Joshua D. and Victor Lavy, "Using Maimonides' Rule to Estimate the Effects of Class Size on Scholastic Achievement," *Quarterly Journal of Economics* 104 (1999), 533-576.

Augurzky, Boris, and Christoph M. Schmidt, "The Evaluation of Community-Based Interventions: A Monte Carlo Study," IZA DP No. 270, March 2001.

Baker, Michael, and Nicole M. Fortin, "Occupational Gender Composition and Wages in Canada, 1987-1988," *Canadian Journal of Economics* 34 (2001), 345-376.

Behrman, Jere R., P. Sengupta, and P. Todd, *Final Report: The Impact of PROGRESA on Achievement Test Scores in the First year*, International Food Policy Research Institute, Food Consumption and Nutrition Division, September 2000.

Bell, Robert M., and Daniel F. McCaffrey, "Bias Reduction in Standard Errors for Linear regression with Multi-stage Samples," Forthcoming, *Survey Methodology* 28 (2002).

Card, David, and Thomas Lemieux, "Dropout and Enrollment Trends in the Postwar Period: What Went Wrong in the 1970s?," NBER Working Paper 7658 (April 2000).

Central Bureau of Statistics, *Statistical Abstract of Israel 53*, Jerusalem: Central Bureau of Statistics, 2002.

Chesher, Andrew, and I. Jewitt, "The Bias of a Heteroscedasticity-Consistent Covariance Matrix Estimator," *Econometrica* 55 (1987), 1217-1222.

Cornfeld, J., "Randomization by Group: A Formal Analysis," *American Journal of Epidemiology* 198 (1978), 100-2.

Dearden, Lorraine, C. Emmerson, C. Frayne, A. Goodman, H. Ichimura, and C. Meghir, *Education Maintenance Allowance: The First year, A Quantitative Evaluation*, Department for Education and Evaluation Research Report RR257, May 2001.

Diehr, Paula, Donald C. Martin, Thomas Koepsell, and Allen Cheadle, "Breaking the Matches in a Paired t-Test for Community Interventions when the Number of Pairs is Small," *Statistics in Medicine* 14 (1995), 1491-1504.

Donald, Stephen, and Kevin Lang (2001), "Inference with Differences-in-Differences and Other Panel Data," Boston University Department of Economics, mimeo, March 2001.

Donner, Allan, K. S. Brown, and P. Brasher, "A Methodological Review of Non-Therapeutic Intervention Trials Employing Cluster Randomization 1979-89," *International Journal of Epidemiology* 19 (1990), 795-800.

Duflo, Esther, and E. Saez, "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," NBER Working Paper 8885, April 2002.

Dynarski, Mark, and Philip Gleason, *How Can We Help? What Have We Learned from Evaluations of Federal Dropout-Prevention Program*, Princeton, NJ: Mathematica Policy research Research report 8014-140, June, 1998.

Eckstein, Zvi, and K.I. Wolpin, "Why Youths Drop Out of High School: The Impact of preferences, Opportunities, and Abilities," *Econometrica* 67 (November 1999), 1295-1340.

Feng, Ziding, P. Diehr, A. Peterson, and D. McLerran, "Selected Statistucal issues in Group Randomized Trials," *Annual Review of Public Health* 22 (2001), 167-87.

Fuller, W.C., C.F. Manski, and D.A. Wise, "New Evidence on the Economic Determinants of Postsecondary Schooling," *The Journal of Human Resources* 17 (Autumn, 1982), 477-498.

Gail, M.H., S. Mark, R. Carroll, S. Green, and D. Pee, "On Design Consideratons and Randomization-based Inference for Community Intervention Trials," *Statistics in Medicine* 15 (1996), 1069-1092.

Gruber, J., "Risky Behavior Among Youths: An Economic Analysis," NBER Working Paper 7781 (July 2000).

Hahn, Andrew, T. Leavitt, and P. Aaron, "Evaluation of the Quantum Opportunities Program (QOP): Did the Program Work?, Heller Graduate School, Center for Human Resources, Brandeis University, June 1994.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62 (1994), 467-75.

Kane, Thamas J., and Douglas O. Staiger, "Improving School Accountability Measures," NBER Working Paper 8156, March 2001.

Keane, Michael P., and K.I. Wolpin, "Eliminating Race Differences in School Attainment and Labor Market Success," *Journal of Labor Economics* 18 (October 2000), 614-52.

Israel Ministry of Education, *Bagrut Test Data 2000*, Jerusalem: Ministry of Education Chief Scientist's Office, April 2001.

Israel Ministry of Education, *The Bagrut 2001 Program: An Evaluation*, Jerusalem: Ministry of Education Evaluation Division, May 2002.

Lavy, Victor, "Evaluating the Effect of Teachers' Group Incentives on Student Achievement," *Journal of Political Economy* 110 (2002), 1286-1317.

Liang, Kung-yee, and Scott L. Zeger, "Longitudinal Data Analysis Using Gerealized Linear Models," *Biometrika* 73 (1986), 13-22.

Long, David M, J.M. Gueron, R.G.Wood, R. Fisher, and V. Fellerath, *LEAP: Three-year Impacts of Ohio's Welfare Initiative to Improve School Attendance Among Teenage Parents*, New York: MDRC, April 1996.

MacKinnon, J.G., and H. White, "Some Heteroscedasticity-Consistent Covariance Matrix Estimators with Improved Finite-Sample Properties," *Journal of Econometrics* 29 (1985), 305-325.

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

Reich, Robert, Op. Ed., *The New York Times* (January 9, 1998).

Rosenbaum, Paul R., *Observational Studies*, New York: Springer-Verlag, 1995.

Schultz, T. Paul, "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program," Economic Growth Center Dsicussion Paper 834, Yale University, August 2001.

Thornquist, Mark D., and G.L. Anderson, "Small-Sample Properties of Generalized Estimating Equations in Group-Randomized Designs with Gaussian Response," Fred Hutchinson Cancer Research Center, Technical Report, 1992.

APPENDIX: ACHIEVEMENT AWARDS PROGRAM RULES AND TIMING

Program Rules

1. Award schedule

Grade	Milestone	Reward (NIS)
10	Tested for at least 1 unit; enrolled in 11th grade	500
	Passed this test	1500
11	Tested for at least 3 units; enrolled in 12th grade	500
	Passed this/these test(s)	1500
12	Completed 14 credit units	1000
	Completed 20 credit units and awarded Bagrut	5000

2. Tests are considered to have been passed if the external component is passed.

3. Only tests in required subjects are eligible for intermediate awards. At the time this program was introduced (January 2001), the required subjects were Bible (2 units), literature (2 units), history (2 units), civics (2 units), composition (2 units), english (3 units), mathematics (3 units). The remaining 5 units can be in any Bagrut-eligible elective subject. Many students, e.g. those competing for admission to selective universities, obtain more than the minimum number of credit units.

4. Awards for achievement in a given year are to be paid in the following school year.

5. All students in treatment schools are eligible.

6. Students with at least 14 units have two chances to take Bagrut exams in 12th grade. Awards will be given to those who pass on the first, second, or any subsequent try.

PROGRAM AND DATA COLLECTION TIME LINE

Program	Schools selected and principals informed	Orientation for principals and students informed	Baseline data	Media coverage	Bagrut tests	Student survey	Re-test	Re-test
School randomization	December 2000	January 2001	January 2001	May 2001	June 2001	Aug/Sept/ Oct 2001	Aug-Sept 2001 (math and english)	December 2001 - January 2002

Note: in March 2001 principals were interviewed to determine whether the program was publicized in schools.

Bonuses were paid in May 2002.

Program	Schools selected and principals informed	Orientation for principals and students informed	Baseline data	Media coverage	Bagrut tests	Student survey	Re-test	Re-test
Student randomization	March 2000	March 2000 (principals and students informed in writing)	March 2000	September 2000	June 2000	July/Aug/ Sept 2000	--	December 2000 - January 2001

Note: in May 2000 principals were interviewed to determine whether the program was publicized in schools.

Bonuses for the June test were first paid in November 2000. Additional bonuses were paid in 2001. Payments were covered by a private donor.

Table 1: Experimental Design for the Pilot Demonstration

Range for p_{1i}	Range for p_{2i}	Threshold for p_{2i}		Offered Bonus		Row Totals
		Low $q_{.22}$	High $q_{.7}$	No	Yes	
		A. All-Treated Sample ($p_{1i}<.053$)				
$[0, q_{.15}]$		--	--	0	146	
		B. Eligible Sample ($.053<p_{1i}<.67$)				
$[q_{.15}, q_{.53}]$	$[0, q_{.22}]$	59	64	0	123	123
	$[q_{.22}, q_{.7}]$	127	125	127	125	252
	$[q_{.7}, 1]$	56	58	114	0	114
	Column Totals	241	248	242	247	489
		C. No-treated Sample ($p_{1i}>.67$)				
$[q_{.53}, 1]$		--	--	612	0	

Note: The table describes the experimental design used in the pilot demonstration. The sample of 1247 described in the table is reduced from the full sample of 1302 because of missing covariates. The notation q_x refers to the x-quantile of the distribution of fitted values, p_{1i} or p_{2i}. Quantiles are school specific, so that about half of the eligible students in each school where offered a bonus.

Table 2: Descriptive Statistics for the Pilot Experiment

Variables by type	Eligible sample (N=489)		All-treated (N=146)	No-treated (N=612)
	Mean (1)	Contrast by High-Low Z_i (2)	Mean (3)	Mean (4)
Program Variables				
Received bagrut	0.329	-0.003	0.027	0.773
	(0.470)	(0.043)	(0.164)	(0.419)
Treated (offered an award)	0.507	0.521		
	(0.500)	(0.039)		
High/Low Threshold (Z_i)	0.505			
	(0.500)			
Eligibility Determination				
Tests Passed at baseline	3.603	0.016	0.432	6.995
	(1.908)	(0.173)	(0.609)	(1.435)
Maximum Score at baseline	79.632	1.243	33.219	96.510
	(8.287)	(0.748)	(26.842)	(3.988)
p_{1i} (fitted value)	0.357	0.021	0.017	0.828
	(0.180)	(0.016)	(0.019)	(0.069)
Treatment Assignment				
Siblings	3.730	-0.227	4.893	4.210
	(2.306)	(0.209)	(2.983)	(2.333)
Father's Education	10.027	0.246	7.949	11.674
	(3.684)	(0.333)	(4.580)	(3.448)
p_{2i} (fitted value)	0.341	0.007		
	(0.082)	(0.007)		
Other Covariates				
Sex	0.409	-0.041	0.500	0.371
	(0.492)	(0.045)	(0.502)	(0.483)
Mother's Education	10.069	0.229	7.391	11.513
	(3.823)	(0.346)	(4.765)	(3.168)
Immigrant	0.139	0.046	0.068	0.141
	(0.346)	(0.031)	(0.253)	(0.348)
Bedouin School	0.102	-0.035	0.349	0.033
	(0.303)	(0.027)	(0.478)	(0.178)

Notes: The table reports descriptive statistics for three subsamples from the pilot experiment conducted in Southern Israel. All 146 students with a very low predicted probability of obtaining a Bagrut (p_{1i}) were offered awards (all-treated sample). None of the 612 students with a high predicted probability of obtaining a Bagrut were offered awards (no-treated sample). Offers in the middle group, referred to as the eligible sample, were determined by whether an index of students' socioeconomic background (p_{2i}) fell below a randomly assigned threshold determined by Z_i. Standard deviations are reported in parentheses in columns of means. Standard errors are reported in parentheses in column 2.

Table 3: Reduced Form Effects in the Pilot Experiment (Eligible Sample)

Dependent Variable	All Eligible Pupils				Jewish Eligible Pupils	
	No Covariates	School Covs p_{2i}	School f.e. p_{2i}	p_{1i}, sex, School f.e., P_{2i}	No Covariates	School f.e. p_{2i}
	(1)	(2)	(3)	(4)	(5)	(6)
Offered	0.521	0.531	0.535	0.535	0.503	0.526
	(0.039)	(0.030)	(0.028)	(0.028)	(0.041)	(0.030)
Received Bagrut	-0.003	0.005	0.001	-0.017	0.013	0.014
	(0.043)	(0.042)	(0.042)	(0.039)	(0.045)	(0.044)

Notes: The table reports coefficients on Z_i (high/low threshold) in regressions with the covariates indicated. The sample size is 489. Columns 4 and 6 report results from models with school fixed effects. School clustering is ignored in columns 1,2, & 5. School covariates consist of a dummy for religious schools and a dummy for the single Bedouin School.

Table 4: Results by Sex in the Pilot Experiment

Dependent Variable	All eligibles		Random-assignment Sample		No-first-stage Sample	
	Boys	Girls	Boys	Girls	Boys	Girls
	(1)	(2)	(3)	(4)	(5)	(6)
Offered Bonus	0.514	0.540	1	1	0.047	0.057
	(0.046)	(0.037)			(0.056)	(0.053)
Received Bagrut	-0.149	0.118	-0.130	0.080	-0.175	0.133
	(0.063)	(0.056)	(0.097)	(0.078)	(0.089)	(0.085)
N	200	289	104	148	96	141

Notes: The table reports coefficients on Z_i (high/low threshold) in regressions using the samples indicated. All models contain p_{2ii} and school fixed effects, as in column 3 in the previous table. The random-assignment sample consists of eligible students with p_{2i} in the interval $[q_{.22}, q_{.7}]$, where the offer of a bonus equals the randomly assigned threshold. The no-first-stage sample has p_{2ii} outside this range, where the offer of a bonus is unrelated to the randomly assigned

Table 5: Descriptive Statistics for the Schools Experiment

Pair	Treated	Non-Complier	Arab School	Religious School	All Pupils						Percent of Students on the Bagrut Track		
					Enrollment			Bagrut Passing Rate					
					1999	2000	2001	1999	2000	2001	1999	2000	2001
1				X	153	173	175	0.046	0	0.091	0.889	0.850	0.914
1	X			X	56	59	45	0.036	0.05	0	0.464	0.949	0.800
2				X	242	169	147	0.054	0.101	0.184	0.083	0.385	0.231
2	X				179	184	145	0.05	0.109	0.11	0.704	0.679	0.676
3					88	99	72	0.114	0	0.056	0.625	0.556	0.750
3	X		X		123	128	99	0.098	0.055	0.03	0.984	0.945	0.919
4					81	68	73	0.148	0.162	0.082	0.926	0.956	0.932
4	X		X		187	221	248	0.134	0.394	0.339	0.738	0.928	0.899
5					125	124	96	0.152	0.105	0.083	0.960	0.952	0.958
5	X			X	55	39	39	0.145	0.077	0.692	0.182	0.410	0.718
6	X				117	123	123	0.171	0.138	0.154	0.530	0.504	0.496
7				X	16	28	16	0.188	0.214	0.375	1.000	1.000	1.000
7	X			X	67	85	58	0.179	0.165	0.483	0.791	0.588	0.793
8				X	57	48	61	0.193	0.771	0.328	0.526	1.000	1.000
8	X				90	96	113	0.189	0.188	0.168	0.744	0.990	0.991
9					61	40	59	0.197	0.35	0	0.344	0.500	0.576
9	X			X	10	14	9	0.2	0.071	0.667	1.000	1.000	1.000
10				X	34	39	26	0.206	0.41	0.654	0.941	1.000	1.000
10	X	X			135	135	108	0.207	0.267	0.361	0.785	0.785	0.769
11					136	148	134	0.213	0.176	0.164	1.000	0.980	0.963
11	X				129	158	152	0.209	0.165	0.092	0.915	1.000	1.000
12			X		19	24	20	0.211	0.667	0.25	1.000	1.000	1.000
12	X			X	32	44	24	0.219	0.25	0.5	1.000	1.000	0.958
13					146	119	123	0.219	0.16	0.211	0.548	0.563	0.593
13	X				85	79	86	0.224	0.367	0.372	0.682	0.785	0.953
14					208	169	186	0.236	0.154	0.274	0.981	0.964	0.984
14	X	X	X		75	50	64	0.227	0.56	0.484	0.907	0.980	0.984
15				X	156	152	163	0.244	0.177	0.331	0.628	0.776	0.939
15	X	X			138	141	152	0.254	0.61	0.467	0.739	0.759	0.618
16				X	102	115	108	0.255	0.226	0.213	0.471	0.809	0.537
16	X				74	60	75	0.257	0.1	0.107	0.784	0.833	0.573
17				X	23	14	16	0.261	0.071	0	0.696	0.857	0.813
17	X		X		76	68	67	0.263	0.441	0.448	1.000	1.000	1.000
18				X	216	209	219	0.273	0.311	0.301	0.958	0.990	0.932
18	X	X			200	148	110	0.275	0.162	0.173	0.680	0.622	0.509
19					141	111	77	0.284	0.54	0.636	0.865	0.892	1.000
19	X	X			123	40	62	0.276	0.025	0.081	0.805	0.975	0.903
20					185	159	111	0.286	0.164	0.126	0.962	0.987	0.973
20	X		X		144	141	167	0.285	0.397	0.353	0.743	0.922	0.731

Notes: The table reports statistics for each school in the 2001 school-level experiment. The control school in pair 6 closed before treatment assignments were announced. Non-compliant schools are treated schools that did not participate in the program.

Table 6: Grouped Estimates for the Schools Experiment

Sample	Mean	Unweighted			Weighted		
		No controls	Sch Cov	Sch Cov + Pair	No controls	Sch Cov	Sch Cov + Pair
		(1)	(2)	(3)	(4)	(5)	(6)
A. 2001 Sample							
1. All Pairs (39 Schools; 3828 Pupils)	0.245	0.075	0.078	0.082	0.048	0.057	0.056
		(0.063)	(0.059)	(0.059)	(0.050)	(0.049)	(0.050)
		[0.062]	[0.057]	[0.038]	[0.047]	[0.047]	[0.033]
2. Balanced Pairs (31 Schools; 2950 Pupils)	0.216	0.119	0.110	0.108	0.083	0.089	0.061
		(0.070)	(0.062)	(0.053)	(0.052)	(0.047)	(0.043)
		[0.067]	[0.057]	[0.034]	[0.048]	[0.042]	[0.028]
3. Low-rate Pairs (28 Schools; 2664 Pupils)	0.222	0.098	0.079	0.087	0.057	0.063	0.066
		(0.076)	(0.065)	(0.048)	(0.057)	(0.055)	(0.046)
		[0.073]	[0.059]	[0.031]	[0.053]	[0.052]	[0.038]
B. 2000 Sample							
1. All Pairs (39 Schools; 4021 Pupils)	0.226	-0.021	-0.019	-0.016	0.048	0.05	0.042
		(0.062)	(0.062)	(0.066)	(0.054)	(0.055)	(0.061)
		[0.061]	[0.059]	[0.043]	[0.054]	[0.052]	[0.039]
2. Balanced Pairs (31 Schools; 3214 Pupils)	0.195	-0.007	-0.004	-0.004	0.052	0.053	0.021
		(0.055)	(0.052)	(0.039)	(0.047)	(0.045)	(0.030)
		[0.054]	[0.049]	[0.025]	[0.049]	[0.044]	[0.019]
3. Low-rate Pairs (28 Schools; 2815 Pupils)	0.188	-0.042	-0.046	-0.046	0.049	0.049	0.010
		(0.073)	(0.074)	(0.066)	(0.058)	(0.059)	(0.059)
		[0.071]	[0.071]	[0.046]	[0.053]	[0.051]	[0.038]

Notes: The table reports treatment effects estimated using school averages. Weighted estimates are weighted by school size. Conventional standard errors are reported in parentheses. Standard errors in brackets are robust (heteroscedasticity consistent).

Table 7: Estimates Using Micro Data for the Schools Experiment

	Two-Step Procedure						Micro Data	
	Unweighted			Weighted				
	No Controls	Sch Cov	Sch Cov + Pair	No Controls	Sch Cov	Sch Cov + Pair	Sch Cov	Sch Cov + Pair
Sample	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				A. 2001 Sample				
1. All Pairs	0.107	0.110	0.116	0.046	0.056	0.068	0.056	0.068
	(0.054)	(0.049)	(0.051)	(0.041)	(0.038)	(0.041)	(0.036)	(0.026)
	[0.053]	[0.047]	[0.034]	[0.038]	[0.036]	[0.026]	{0.039}	{0.038}
2. Balanced Pairs	0.145	0.136	0.138	0.075	0.082	0.074	0.081	0.073
	(0.062)	(0.052)	(0.050)	(0.045)	(0.036)	(0.038)	(0.032)	(0.024)
	[0.059]	[0.049]	[0.032]	[0.041]	[0.032]	[0.025]	{0.034}	{0.035}
3. Low-rate Pairs	0.137	0.120	0.133	0.051	0.056	0.083	0.055	0.089
	(0.067)	(0.055)	(0.046)	(0.049)	(0.044)	(0.043)	(0.040)	(0.036)
	[0.064]	[0.050]	[0.031]	[0.044]	[0.041]	[0.032]	{0.044}	{0.053}
				B. 2000 Sample				
1. All Pairs	-0.009	-0.007	-0.003	0.028	0.032	0.039	0.031	0.040
	(0.050)	(0.049)	(0.051)	(0.043)	(0.043)	(0.047)	(0.041)	(0.030)
	[0.049]	[0.047]	[0.033]	[0.043]	[0.042]	[0.031]	{0.044}	{0.045}
2. Balanced Pairs	0.015	0.017	0.017	0.042	0.044	0.033	0.043	0.033
	(0.044)	(0.040)	(0.034)	(0.038)	(0.034)	(0.029)	(0.033)	(0.019)
	[0.043]	[0.038]	[0.022]	[0.040]	[0.033]	[0.019]	{0.037}	{0.028}
3. Low-rate Pairs	-0.023	-0.028	-0.023	0.034	0.035	0.016	0.033	0.009
	(0.057)	(0.056)	(0.048)	(0.044)	(0.043)	(0.043)	(0.035)	(0.029)
	[0.055]	[0.053]	[0.034]	[0.038]	[0.036]	[0.031]	{0.039}	{0.046}

Notes: Columns 1-6 report estimates using school fixed effects from a student-level regression included lagged score quartiles. Conventional standard errors are shown in parentheses. Standard errors in brackets are robust (heteroscedasticity-consistent). Columns 7 and 8 report regression results using micro data, with controls for lagged score quartiles. Standard errors in parentheses are adjusted for school clustering using the formulas in Liang and Zeger (1986). Standard errors in braces use MacCaffrey and Bell's (2002) BRL estimator.

Table 8: Effects on Early and Late Bagrut Rates by Quartile of Previous Test Scores

		Estimates by quartile: June 2001				Estimates by quartile: Winter 2002			
		1st quartile	2nd quartile	3rd quartile	4th quartile	1st quartile	2nd quartile	3rd quartile	4th quartile
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. 2001 Sample									
1. All Pairs	*Treatment*	0.037	0.016	0.106	0.115	0.034	0.013	0.126	0.071
(mean=.245)	*Effects*	(0.044)	(0.031)	(0.041)	(0.063)	(0.044)	(0.035)	(0.042)	(0.062)
		{0.049}	{0.041}	{0.051}	{0.078}	{0.046}	{0.044}	{0.053}	{0.078}
	Control group means	0.010	0.092	0.292	0.486	0.010	0.123	0.343	0.570
2. Balanced Pairs	*Treatment*	0.041	-0.026	0.096	0.182	0.032	-0.015	0.126	0.137
(mean=0.216)	*Effects*	(0.058)	(0.028)	(0.042)	(0.057)	(0.054)	(0.031)	(0.042)	(0.052)
		{0.067}	{0.038}	{0.053}	{0.066}	{0.058}	{0.040}	{0.052}	{0.062}
	Control group means	0.005	0.063	0.228	0.406	0.005	0.087	0.269	0.499
3. Low-rate Pairs	*Treatment*	0.071	0.000	0.132	0.159	0.067	0.003	0.158	0.118
(mean=.222)	*Effects*	(0.056)	(0.041)	(0.053)	(0.068)	(0.043)	(0.041)	(0.049)	(0.064)
		{0.076}	{0.058}	{0.069}	{0.080}	{0.052}	{0.054}	{0.063}	{0.076}
	Control group means	0.011	0.096	0.246	0.435	0.011	0.123	0.293	0.544
B. 2000 Sample									
1. All Pairs	*Treatment*	0.027	0.012	0.056	0.065	0.037	0.019	0.051	0.055
(mean=226)	*Effects*	(0.040)	(0.033)	(0.049)	(0.061)	(0.041)	(0.033)	(0.049)	(0.056)
		{0.041}	{0.046}	{0.064}	{0.079}	{0.044}	{0.044}	{0.061}	{0.072}
	Control group means	0.019	0.086	0.24	0.498	0.028	0.119	0.287	0.561
2. Balanced Pairs	*Treatment*	0.026	0.004	0.048	0.055	0.032	0.009	0.035	0.047
(mean=.195)	*Effects*	(0.041)	(0.023)	(0.045)	(0.056)	(0.043)	(0.026)	(0.043)	(0.058)
		{0.044}	{0.031}	{0.055}	{0.064}	{0.046}	{0.032}	{0.051}	{0.066}
	Control group means	0.002	0.067	0.192	0.425	0.032	0.101	0.256	0.530
3. Low-rate Pairs	*Treatment*	0.017	-0.023	-0.004	0.039	0.030	-0.006	0.016	0.053
(mean=.188)	*Effects*	(0.045)	(0.032)	(0.049)	(0.067)	(0.046)	(0.039)	(0.054)	(0.066)
		{0.053}	{0.044}	{0.061}	{0.087}	{0.055}	{0.051}	{0.065}	{0.080}
	Control group means	0.000	0.056	0.209	0.413	0.005	0.086	0.242	0.455

Notes: The table reports estimated treatment effects for early and late Bagrut outcomes. Treatment effects vary by quartile of summary Bagrut scores through January 2001 or January 2000. Standard errors in parentheses are adjusted for clustering using formulas in Liang and Zeger (1986) and in braces using MacCaffrey and Bell's (2002) BRL estimator. The models correspond to those in column 8 of Table 7 (control for school covariates and pair effects).
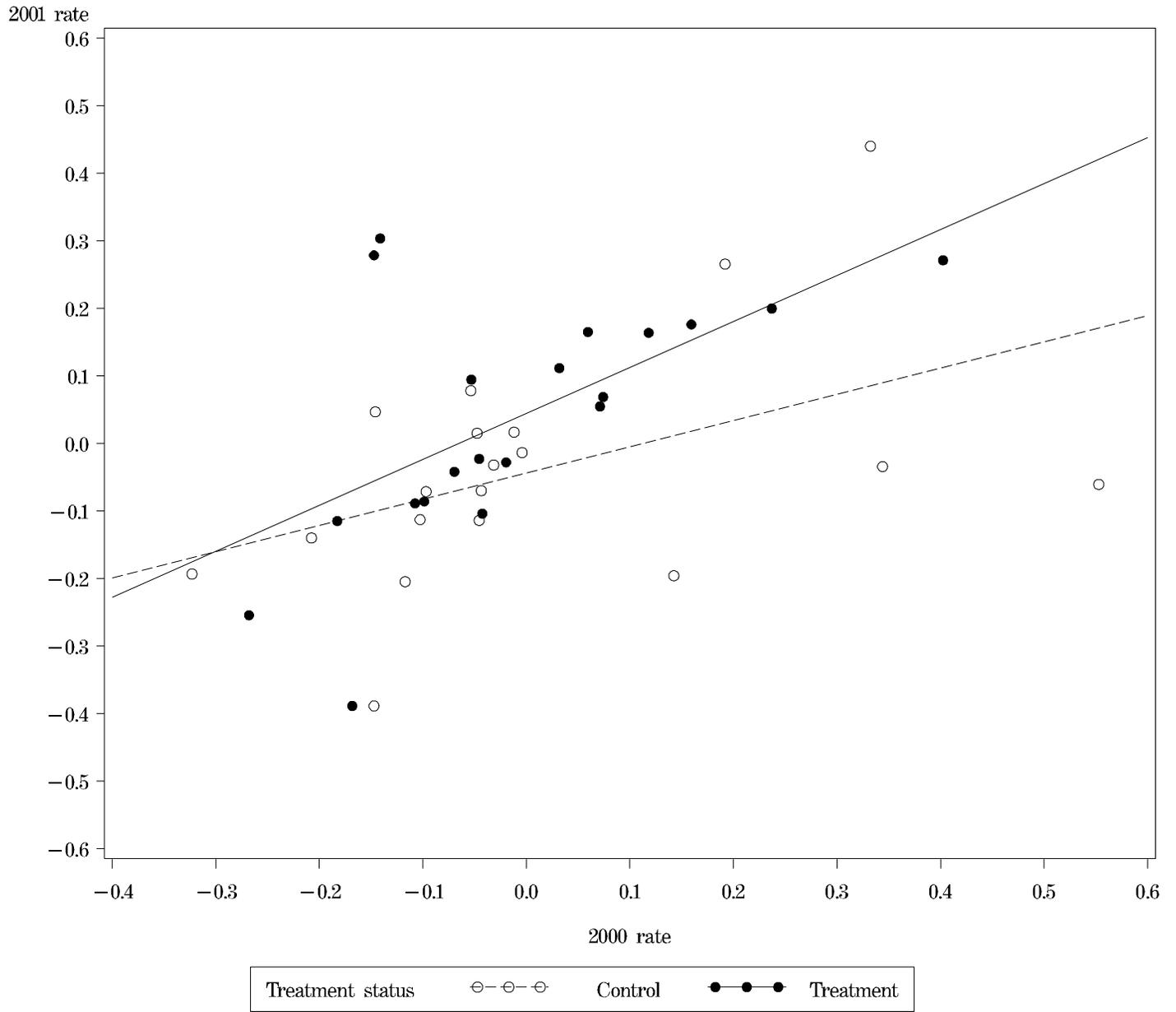
Figure 1. 2001 Bagrut rate vs. 2000 Bagrut rate by treatment status.
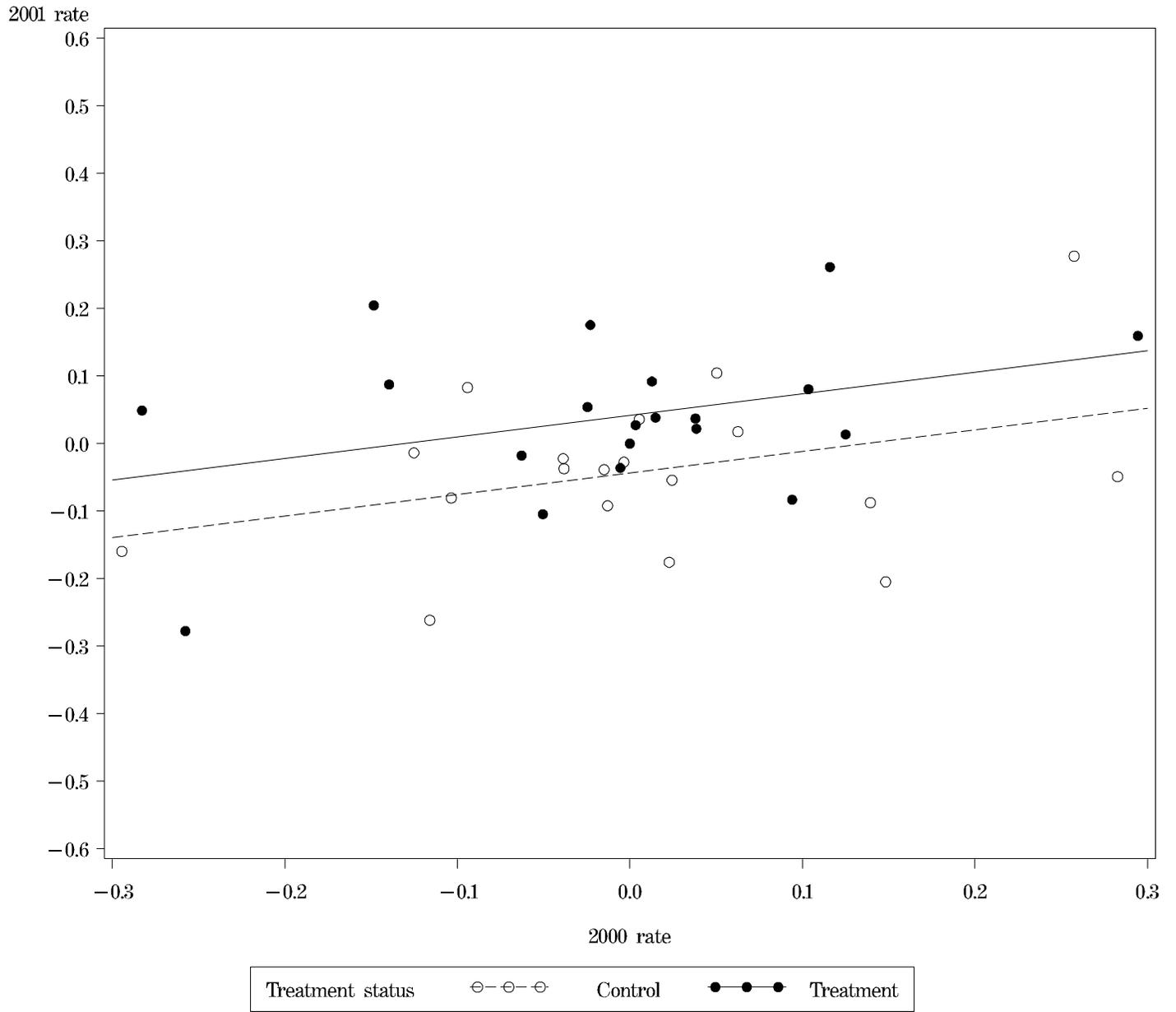Residuals from regressions on school covariates.

Figure 2. 2001 Bagrut rate vs. 2000 Bagrut rate by treatment status.
Residuals from regressions on school covariates and pair effects.