DISCUSSION PAPER SERIES

IZA DP No. 1146

The Effect of High Stakes High School Achievement Awards: Evidence from a School-Centered Randomized Trial

Joshua D. Angrist
Victor Lavy

May 2004

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Effect of High Stakes High School Achievement Awards: Evidence from a School-Centered Randomized Trial

Joshua D. Angrist
Massachusetts Institute of Technology, NBER and IZA Bonn

Victor Lavy
Hebrew University of Jerusalem

Discussion Paper No. 1146
May 2004

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 1146
May 2004

ABSTRACT

The Effect of High Stakes High School Achievement Awards: Evidence from a School-Centered Randomized Trial*

In many countries, college-bound high school seniors must pass a test or series of tests. In Israel, this requirement is known as the "Bagrut", or matriculation certificate, obtained by passing a series of subject tests. In spite of the Bagrut's value, Israeli society is marked by vast differences in Bagrut rates by region and socioeconomic status. We attempted to increase the likelihood of Bagrut certification among low-achieving students by offering substantial cash incentives to high school seniors in an experimental demonstration program. As a theoretical matter, such 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. The experiment studied here used a school-based randomization design offering awards to all students in treated schools who passed their exams. Randomization was imperfect because of the clustered design. We discuss alternative strategies for dealing with clustering in research of this type. On balance, the estimates point to a substantial and statistically significant treatment effect for students close to the margin for certification. We also look at a number of mediating outcomes in an effort to determine how students responded to incentives. These results show students took more tests and were more likely to accumulate the number of credit units required for Bagrut success.

JEL Classification: I21, I28, C93

Keywords: performance incentives, school reform, clustering

Corresponding author:

Joshua D. Angrist
Department of Economics
MIT
50 Memorial Drive
Cambridge, MA 021242-1347
USA
Email: angrist@mit.edu

* Special thanks go to Rema Hanna and Alex Levkov for outstanding research assistance in Cambridge and Jerusalem, and to Dan McCaffrey for sharing his BRL code. Thanks also to Daron Acemoglu, Abhijit Banerjee, David Card, Sue Dynarksi, Ron Ehrenberg, Jinyong Hahn, Guido Imbens, Alan Krueger, Adriana Kugler, Kevin Lang, Thomas Lemieux, and seminar participants at Berkeley, Boston University, Case Western, CEMFI, Harvard/MIT, Hebrew University, McMaster, Princeton, SOLE, Tel Aviv, UCLA/RAND, University of Colorado, and Washington University for helpful discussions and comments. The 2001 Achievement Awards program was funded by the Israel Ministry of Education and administered by the division for secondary schools. We also gratefully acknowledge funding under NIH grant 1R01HD043809-01A1. The statements in the paper reflect the views of the authors and have not been endorsed by program sponsors or funding agencies. This is a substantially revised version of NBER Working paper 9389.

I. Introduction

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, the high school matriculation certificate or Bagrut, awarded after a sequence of subject tests, 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, with no apparent effect. This echoes similar findings from randomized trials in the U.S., where an array of service-oriented anti-dropout demonstrations for American teens failed to increase 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 immediate financial rewards may tip the scales in favor of schoolwork. In this paper, we report on an experimental demonstration project that provided cash awards for low-achieving high school students in Israel. The intervention 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 raises a number of practical and research-design questions such as how the incentives should be structured, the appropriate experimental design, and whether incentives that target entire schools are more likely to be effective than those that target individual students. Based on preliminary results from a smaller pilot study detailed in our working paper (Angrist and Lavy, 2002), we settled on a school-based demonstration. The pilot project, which had no effect on achievement, involved random assignment of students within schools. In contrast, the main experimental intervention

assigned treatment to entire schools and was implemented with the cooperation of principals and administrators in treated schools. This intervention offered modest awards for completion of individual subject tests and for continued school enrollment, with the highest awards reserved for 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 inference.

Though the intervention studied here is unusual (and was indeed controversial in Israel), there is growing interest in student incentive programs in secondary education. Other interventions in this spirit include the Quantum Opportunities Program (Maxfield, Schirm, and Rodriguez-Planas, 2003); 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, 2004); a program in Colombia (PACES) that provided private school vouchers (Angrist, Bettinger, Bloom, King, and Kremer, 2002); and a recent randomized demonstration of a scholarship program for girls in Kenya (Kremer, Miguel, and Thornton, 2003).[2] The

[2]QOP combined services for children in AFDC families with modest financial incentives for enrollment in a small randomized study. 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-

Achievement Awards demonstration 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. Finally, any merit-based scholarship, such as the long-running but little-studied National-Merit and National-Achievement Scholarship Programs, have elements in common with Achievement Awards.[3]

The rest of the paper is organized as follows. Section II sketches some of the theoretical background motivating our intervention. Section III describes the demonstration program in detail and Section IV presents descriptive statistics and outlines the econometric framework. Section V discusses the results, 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 pre-intervention achievement levels that put them in a position to benefit from additional effort. We further substantiate the causal interpretation of our results with a differences-in-differences type analysis that controls for imbalances in baseline outcomes, and briefly explore the proximate channels through which awards appear to have increased certification rates. These results suggest winners used a more ambitious test-taking strategy and had a higher probability of success, at least on the margin. Finally, Section VI concludes and discusses directions for further work.

II. Theoretical Context

Why do young men and women fail to complete high school? Why don't more people go to college?

---

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 and the study by Kremer, Miguel, and Thornton (2003) are the first demonstration projects to evaluate substantial achievement-based payments to students using a randomized experimental design.

[3]The National Merit programs give recognition and modest cash awards to a handful of high-achieving students based on their PSAT scores. As an interesting methodological note, the regression-discontinuity research design was introduced by Campbell (1969) as a strategy to evaluate the effect of these programs.

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 and social 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 participation in widely-available 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 of the students who fail to complete a Bagrut still finish their secondary schooling. Nevertheless, as for 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.[4]

The Achievement Awards demonstration was 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 for students 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.

III. Program Details

A. 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 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.[5]

---

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

[5]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 (i.e., anonymously graded) 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.

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.

B. Implementation and Research Design

In December 2000, we selected 40 high schools with very 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.[6] 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). Moreover, 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.[7] 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.

To kick off the school-based demonstration, an orientation with principals and administrators from the 20 treatment schools was held in January 2001. Some principals chose not to participate, though most

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

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

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. The award schedule is detailed in the appendix. The program was meant to last 3 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 deviation from the intended scenario, it is 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

[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.

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 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]

IV. Descriptive Statistics and Econometric Framework

A. Description of School Means

Table 1 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. The principals of three non-compliant schools had taken no concrete actions to inform students or teachers about the program and/or indicated that they did not wish to participate. School administrators in two other schools designated as non-compliant hoped to participate but submitted student rosters shortly after the deadline.

The enrollment figures in Table 1 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 discussed above, this is much lower than the

[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.

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. For example, Kane and Staiger (2002) report that much of the year-to-year variation in school performance in North Carolina is due to school-level random shocks that come from sources other than sampling variance and permanent differences in school characteristics.

Table 1 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 (the other track is vocational). 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 potentially confounding factors, even after matching. This is

especially relevant in view of the unstable Bagrut rates in Table 1. For a subset of the analyses that follow, we attempted to improve treatment-control balance by discarding the 4 pairs with the largest standardized differences (as measured by t-statistics) in 2000 (i.e., pre-treatment) Bagrut rates. Other econometric issues and our estimation framework are detailed 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 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, 1991).

A second adjustment for compliance is suggested by the descriptive statistics in Table 1. 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 1 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 some 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 of individual students for both programmatic and logistical reasons. First, school-based assignment reduces the perception of unfairness that may be associated with randomization. Second, students not offered treatment may nevertheless be affected by the treatment received by other students 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.[10]

The most important statistical issue in school-level GRTs is whether to treat groups (schools) or students as the unit of observation for data analysis, 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 are biased and tend to be too low, making measures of precision overly optimistic.

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. An advantage of GEE is that it is very flexible and increasingly available in proprietary software.[11] 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

[10]A recent experiment offering incentives for financial education illustrates this point (Duflo and Saez, 2002).

[11]GEE standard errors are produced by the Stata "Cluster" option and SAS GENMOD procedure.

effects models). GRTs often have too few clusters for asymptotic formulas to provide an acceptably accurate approximation to the finite-sample sampling distribution. As with tradition Moulton-type or design-effect adjustments, GEE standard errors are also biased downwards.[12]

A simple alternative to micro-level analyses is to work with grouped data, which in this case means school averages. With an adequate group size, average Bagrut rates are approximately Normally distributed. T-tests are therefore likely to be valid for a grouped analysis, even with a moderate number of groups (see, e.g., Donald and Lang, 2001). The typical grouped equation in our analysis can be written:

$$\bar{y}_{jt} = \alpha_j + x_{jt}'\beta + \delta Z_{jt} + \eta_{jt}, \quad (1)$$

where j=1, . . ., 20 indexes pairs; t=0,1 indexes treatment status, and Z_{jt} is assigned treatment status. The dependent variable, $\bar{y}_{jt}$, is the school average Bagrut rate, x_{jt} is a vector of school characteristics (dummies for Arab and religious schools), and η_{jt} is group error. Some models also include pair effects, α_j.

Some of the grouped estimates are weighted by the school size, n_{jt}. Classical results on regression efficiency make it seem natural to weight a grouped equation by group size, and weighted estimation using group means produces the same estimates as micro data when there are no micro 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.

While unlikely to be generate misleading inferences, grouped analyses have two drawbacks. First, micro data on student characteristics may reduce the variability in outcomes. The potential benefit from control for covariates is highlighted by the temporal variation in group means evident in Table 1. A student-level regression with individual control variables may absorb some of this variation. Second, 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

[12]See, e.g., Thornquist and Anderson (1992) and Wooldridge (2003).

may therefore be less powerful than those based on micro data.

The model we used to construct micro estimates from data on individual students can be written:

$$y_{ijt} = \alpha_j + x_{jt}'\beta + \sum_q d_{qi}\mu_q + \delta Z_{jt} + \epsilon_{ijt}, \quad (2)$$

where i indexes students; and the j and t subscripts are as before. Also as in equation (1), the grouped covariates include pair effects, α_j, and the two school-level covariates denoted, x_{jt}. 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 an excellent predictor of students' Bagrut status. Other characteristics, such as measures of family background, add little after conditioning on lagged scores.

As noted earlier, inference with micro data must take account of clustering in the error term, ϵ_{ijt}, but traditional cluster adjustments are likely to be misleading. 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 statistical tests of the correct size in traditional random effects models with Normally distributed errors.

In a second approach to the clustering problem with micro data, we 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 (2), and then regressing the estimated fixed effects on treatment status and other school-level covariates in a grouped equation like (1). In particular, we first estimate

$$y_{ijt} = \mu_{jt} + \sum_q d_{qi}\mu_q + \xi_{ijt}, \quad (3a)$$

and then regress $\hat{\mu}_{jt}$, the estimated μ_{jt}, on the same covariates as in equation (1). Thus, the second step is

$$\hat{\mu}_{jt} = \alpha_j + x_{jt}'\beta + \delta Z_{jt} + \nu_{jt} \quad (3b)$$

The two-step procedure provides an appealing compromise between grouped and individual analyses

in that it uses the micro-data to reduce the dispersion in group means, while inference is conservative in the sense that no credit is 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 many 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.[13]

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. Moreover, ignoring stratification variables may 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, 2004). On the other hand, with few pairs, a chance association between pair characteristics and treatment status is possible.

V. Results

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

[13]The GLS version of the two-step procedure was suggested by Amemiya (1978), who shows that it is equivalent to GLS using micro data with a clustered error structure. Hansen (2003) shows that first-step estimation of the fixed effects does not affect the limiting distribution in the second step.

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 2, 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.[14]

The standard errors quoted above (and reported in Table 2 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 for the grouped estimates.

The balanced and low-rate subsamples generate larger treatment effects than the full sample, again

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

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 included dummies for each quartile of the average score distribution. We used quartile dummies instead of, say, linear control for lagged scores, to facilitate an analysis conditional on lagged scores. The quartile dummies are a powerful predictor of students' ultimate Bagrut status: the probability of being awarded a Bagrut 2001 was about 1% in the lowest quartile, 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 by equations (3a) and (3b) 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 3, 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.[15]

The last two columns in Table 3 report treatment effects estimated with micro data. The standard errors reported in parentheses use the 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.

Estimates by Subgroup

On balance, the results in Tables 2 and 3 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

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

where δ_q is a quartile-specific treatment effect and μ_q is a quartile main effect. Students with very low scores

[15]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.

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 4. 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 3. 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 2 and 3 and the first 4 columns of Table 4 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 of the extra round seemed to be to get Bagrut rates up by easing standards. In any case, the last 4 columns of Table 4, 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.[16] 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.

C. Additional Estimates

Differences-in-Differences

While Table 4 shows a clear pattern of significant treatment effects for students in the upper quartiles of the lagged score distribution, with no significant effects in the pre-treatment year, a natural question raised by these results is whether the estimates for 2001 and 2000 are significantly different *from each other*. To compare the estimates for 2001 and 2000 we estimated stacked models for the two years controlling for additive school effects. The coefficient of interest in this specification is the interaction between a dummy for "post", i.e., 2001, and the treatment indicator, Z_{jt}. The estimates are limited to the sample of students with lagged scores in the 3rd and 4th quartile, since Table 4 strongly suggests this is where the action is. The resulting estimates therefore amount to a differences-in-differences (DD) interpretation of the results in columns 3 and 4 of Table 4. To implement this, we use a grouped-data setup analogous to that used for Table 2 since Table 3 shows BRL standard errors with micro data to be essentially the same as those for two-step models. Also, we have conditioned out most of the relevant micro variation by limiting the sample to the upper lagged-score quartiles.

The stacked equation used to estimate DD models can be written as:

$$\bar{y}_{jtqp} = \mu_{jt} + \kappa_q + \pi_p + \beta'(x_{jt}d_p) + \delta(Z_{jt}d_p) + \epsilon_{jtqp}, \qquad (5)$$

where p=0, 1 for pre/post (i.e., 2000 and 2001), d_p=1(p=1), and κ_q and π_p are lagged-score-quartile (for quartile 3 vs. 4) and period effects. The dependent variable, $\bar{y}_{jtqp}$, is the Bagrut rate in school j,t in period p and quartile q. In addition to providing a check on the precision of the 2001-vs.-2000 contrast in treatment

[16]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.

effects, equation (5) can be seen as a framework for the control of omitted school effects that are correlated with treatment status. The validity of this control strategy, however, turns on the validity of an additive conditional mean function as a specification for potential outcomes in the absence of treatment. The additive specification is not entirely natural since the dependent variable is a sample proportion. We therefore report marginal effects from a version of equation (5) that replaces the dependent variable with empirical Logits, i.e., $ln[\bar{y}_{jtqp}/(1-\bar{y}_{jtqp})]$, for cells where $\bar{y}_{jtqp}$ is non-zero.[17]

Unweighted estimates of δ in equation (5), reported in Table 5, are on the order of 15 percentage points in all three samples, regardless of whether Logit or linear models are used for estimation. The standard errors are about 6 percent, even without the heteroscedasticity correction, which probably biases standard error estimates downwards. As in the previous tables, weighted estimates show smaller effects, in this case mostly in the 6-8 percentage point range (with a few considerably larger) and closer to the margin of statistical significance than the corresponding weighted estimates. Still, there is a clear pattern of treatment effects, with significant estimates in column 6 for all three samples. Logit marginal effects tend to be somewhat larger, and more of these are significantly different from zero than the estimates from linear models. Thus, Table 5 supports the interpretation of earlier results as causally related to treatment.

Channels for Improvement

By early Senior year, Israeli high school students have made many commitments regarding their curriculum and, of course, past learning effort. Nevertheless, the results in Tables 2-5 suggest at least some students were able to respond to the Achievement Award incentives program in a manner that paid off. These students were probably in the group identified by the Ministry of Education as marginal, that is, within reach of Bagrut certification even if they fail to achieve it. In our sample, such students had lagged scores that gave them a chance of obtaining a Bagrut. In this last section, we try to understand the channels through which

[17]School covariates and pair effects are entered with time-varying coefficients when they are included in the stacked models.

Achievement Awards produced higher Bagrut rates for this group.

To obtain a Bagrut, students must clear a number of hurdles. These include passing subject tests worth a minimum of 20 credit units, a 2 unit Composition requirement, self-determined credit-unit targets in Math and English, and passing at least one advanced subject test worth 5 units. A natural question for our analysis is whether treated students were more ambitious, i.e., took more exams, or simply did better on exams they would have taken anyway. This is a difficult question to answer since we don't know what exams treated students would have taken had they not been treated. On the other hand, we can assess the impact of treatment on student ambition as measured by credit units *attempted*. For example, the basic Math curriculum, which awards 3 units, is completed by taking two tests, one for a single unit, and one for two more units. Students may have responded to program incentives by taking both tests where they would have previously taken only one (say in 11th grade) or none at all.[18]

The top of Table 6 reports estimated treatment effects on indicators for units-attempted thresholds. In particular, the first outcome is a dummy indicating 18 or more units attempted, followed by indicators for 20, 22, and 24 units attempted. Recall that students must obtain 20 credit units if they are ultimately to be awarded a Bagrut (a necessary though not sufficient condition). We would therefore expect to see some shift in effort at or near the 20 unit threshold.

The results in Table 6 show that students in treated school indeed became more ambitious, where ambition is gauged by units attempted in June 2001. Estimated treatment effects are reported from weighted and unweighted regressions for 2000, for 2001, and from a stacked differences-in-differences-type equation, using the same sample as for Table 5. The unweighted stacked results show a pattern of positive and significant effects on attempted units, peaking at 20 and 22 units as might be expected if the response pattern is due to a more ambitious test-taking strategy for students close to the passing margin. For example, the effect on the probability of obtaining 18-plus units is .044 (s.e.=.041) unweighted, but the unweighted effect

[18]The more advanced biology and physics programs require five different exams, worth one credit unit each.

on obtaining 20-plus units is .098 (s.e.=.044). The corresponding weighted results are .033 (s.e.=.031) and .070 (s.e.=.034). The unweighted estimates also show a significant effect at 22 units, while neither weighted or unweighted show an effect at 24 units.

The increase in units attempted translated into an increase in units awarded. This can be seen in the next four rows of the table, which report indicators for obtaining 18, 20, 22, and 24 units, this time measured as of June 2002.[19] For example, estimates from the unweighted stacked models are .066 (s.e.=.043) for the probability of obtaining 18-plus units, .085 (s.e.=.046) for the probability of obtaining 20-plus units, .074 (s.e.=.044) for the probability of obtaining 22 units, and .048 (s.e.=.044) for the probability of obtaining 24 units. A similar pattern is observed in the weighted stacked results though, as in earlier tables, the weighted estimates are smaller than the unweighted.

The results for units attempted show that at least some and perhaps most of the increase in units awarded is due to an increase in units attempted. An interesting question that we cannot answer directly is whether the increase in units awarded is *solely* due to the increase in units attempted. This is because selection bias contaminates any comparisons of units awarded that are made *conditional* on units attempted. It may be that the students who would have attempted 20 units in either case do somewhat better under treatment, but treatment probably induces somewhat weaker students to take additional Bagrut tests, generating a negative selection bias in conditional-on-attempt comparisons by treatment status. Thus, differences in success rates conditional on having attempted a certain number of units do not have a straightforward causal interpretation.

The remainder of Table 6 shows estimated effects on subject-related distributional requirements and a Bagrut-like composite variable. Distributional outcomes include the probability of passing an advanced subject test, and the Math, English, and Composition requirements. The effects on Math achievement are

[19]Results for units awarded as of June 2001 are similar. Units attempted are measured as of June 2001 because we do not have the data necessary to compute units attempted for the later period. The mean units awarded are close to the mean units attempted (and in one case exceed units attempted) because of the timing discrepancy.

large and significant in some specifications, but overall there is no clear pattern of effects on these outcomes. There is a substantial and significant effect – .070 in the weighted stacked specification – on a composite outcome which pools individual requirements much as is done for the determination of ultimate Bagrut status (with no corresponding effect on the composite in 2000). This suggests the program caused students to act on a number of margins, though the single most important channel appears to be a more ambitious test-taking strategy as reflected in units attempted.

VI. Summary and Conclusions

Although the evidence is not seamless, the results reported here suggest worthwhile gains in matriculation rates can be obtained by offering cash awards to students in low-achieving schools. A causal interpretation of the main results is supported by an analysis in subgroups of students grouped their by lagged test scores, in differences-in-differences estimates, and by an analysis of the channels through which students may have responded to Achievement Awards incentives.

The value of the achievement awards was substantial from the point of view of high school seniors, but probably pales in comparison with the 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 higher wages (not controlling for schooling), 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 in the US appear to have been

ineffective. Not long after the Achievement Awards demonstration, however, the Israeli 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 much, while producing an effect only about 50 percent larger.

Although our study shows that cash incentives can induce actions that improve outcomes on a high stakes test, we have not shown that the incentives offered here improve outcomes by increasing learning. The long-run payoff may come from the fact that the Bagrut is institutionally linked to further schooling and to opportunities in the labor market. It is worth noting, however, that the period between March and June of an Israeli high school student's senior year is almost totally devoted to intense Bagrut preparation for all students who intend to take Bagrut tests. Other students effectively finish high school at the end of March. The Achievement Awards program therefore likely caused some treated students to use this preparation period for studying. Since Bagrut tests are tightly tied to core subject matter, a period of intense study focused on these tests may indeed have a human capital payoff.

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 and are essentially the same as the standard errors from a hybrid two-step procedure. On the other hand, conventional cluster-adjusted standard errors appear to be misleading.

Finally, the analysis here covers the immediate short-run impact on the Achievement Awards programs' target objective, high school matriculation status, and components that determine this outcome.

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. We are also conducting additional research on alternative modes of inference in experiments of this type.

REFERENCES

Amemiya, Takeshi, "A Note on a Random Coefficient Model," *International Economic Review* 19 (1978), 793-796.

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," *The Review of Economics and Statistics* 86 (2004), 58-72.

Angrist, J.D. and Guido Imbens, "Sources of Identifying Information in Evaluation Models," NBER Technical Working Paper 117, December 1991.

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.

Angrist, Joshua D. and Victor Lavy, "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," NBER Working Paper 9389, December 2002.

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," *Survey Methodology* 28 (2002).

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

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 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 Statistical 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 Considerations 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).

Hansen, Christian, "Generalized Least Squares Estimation in Differences-in-Differences and Other Panel Models," MIT department of Economics, mimeo, November 2003.

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.

Kane, Thomas J., and Douglas O. Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives* 16 (Fall 2002), 91-114.

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.

Kremer, Michael, Edward Miguel, and Rebecca Thornton, "Incentives to Learn," Harvard Department of Economics, mimeo, October 2003.

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

Liang, Kung-yee, and Scott L. Zeger, "Longitudinal Data Analysis Using Generalized 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.

Maxfield, Myles, Allen Schirm, and Nuria Rodriguez-Planas, "The Quantum Opportunities Program Demonstration: Implementation and Short-Term Impacts," Mathematica Policy Research Report 8279-093, Washington, DC: Mathematica Policy Research, Inc., August 2003.

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," *Journal of Development Economics* 74 (June), 199-250.

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.

Wooldridge, Jeffrey M., "Cluster-Sample Methods in Applied Econometrics," *American Economic Review* 93 (May 2003), 133-138.

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.

Table 1: Descriptive Statistics for the Schools Experiment

Pair	Treated	Non-Complier	Arab School	Relig. School	All Pupils: Enrollment 1999	Enrollment 2000	Enrollment 2001	Bagrut Passing Rate 1999	Bagrut Passing Rate 2000	Bagrut Passing Rate 2001	Percent of Students on Bagrut Track 1999	Bagrut Track 2000	Bagrut Track 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 2: Grouped Estimates for the Schools Experiment

		Unweighted			Weighted		
Sample	Mean	No controls (1)	Sch Cov (2)	Sch Cov + Pair (3)	No controls (4)	Sch Cov (5)	Sch Cov + Pair (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 3: 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 4: 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	0.010	0.121	0.069
(mean=0.245)	*Effects*	(0.044)	(0.031)	(0.041)	(0.063)	(0.043)	(0.035)	(0.042)	(0.062)
		{0.049}	{0.041}	{0.051}	{0.078}	{0.045}	{0.044}	{0.053}	{0.077}
	Control group means	0.010	0.092	0.292	0.486	0.010	0.125	0.347	0.572
2. Balanced Pairs	*Treatment*	0.041	-0.026	0.097	0.182	0.033	-0.019	0.119	0.136
(mean=0.216)	*Effects*	(0.057)	(0.028)	(0.043)	(0.058)	(0.054)	(0.031)	(0.042)	(0.052)
		{0.067}	{0.038}	{0.053}	{0.066}	{0.058}	{0.040}	{0.051}	{0.061}
	Control group means	0.005	0.063	0.228	0.406	0.005	0.091	0.275	0.501
3. Low-rate Pairs	*Treatment*	0.071	0.000	0.132	0.159	0.067	-0.002	0.148	0.115
(mean=0.222)	*Effects*	(0.056)	(0.041)	(0.053)	(0.068)	(0.043)	(0.041)	(0.049)	(0.063)
		{0.076}	{0.058}	{0.069}	{0.080}	{0.051}	{0.053}	{0.062}	{0.074}
	Control group means	0.011	0.096	0.246	0.435	0.011	0.127	0.299	0.547
		B. 2000 Sample							
1. All Pairs	*Treatment*	0.027	0.012	0.056	0.065	0.037	0.019	0.051	0.050
(mean=0.226)	*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}	{0.043}	{0.044}	{0.061}	{0.071}
	Control group means	0.019	0.086	0.240	0.498	0.028	0.119	0.287	0.565
2. Balanced Pairs	*Treatment*	0.026	0.004	0.048	0.055	0.033	0.009	0.034	0.041
(mean=0.195)	*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}	{0.043}	{0.044}	{0.061}	{0.071}
	Control group means	0.002	0.067	0.192	0.425	0.009	0.092	0.238	0.490
3. Low-rate Pairs	*Treatment*	0.017	-0.023	-0.004	0.039	0.032	-0.006	0.015	0.047
(mean=0.188)	*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	0.460

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 3 (control for school covariates and pair effects).

Table 5: Grouped and Stacked DD Estimates with School Effects

Sample	Mean	Unweighted: No Controls (1)	Unweighted: School Covs (2)	Unweighted: Sch Cov + Pair (3)	Weighted: No Controls (4)	Weighted: School Covs (5)	Weighted: Sch Cov + Pair (6)
A. Linear Models							
1. All Pairs (154 school-year averages)	0.418	0.151	0.155	0.158	0.061	0.065	0.078
		(0.060)	(0.058)	(0.056)	(0.044)	(0.044)	0.043
		[0.052]	[0.050]	[0.046]	[0.038]	[0.038]	[0.033]
2. Balanced Pairs (122 school-year averages)	0.379	0.157	0.142	0.146	0.078	0.074	0.103
		(0.070)	(0.064)	(0.064)	(0.049)	(0.045)	(0.047)
		[0.059]	[0.054]	[0.052]	[0.039]	[0.036]	[0.034]
3. Low-rate Pairs (110 school-year averages)	0.372	0.189	0.170	0.192	0.061	0.064	0.179
		(0.078)	(0.075)	(0.073)	(0.056)	(0.057)	(0.057)
		[0.067]	[0.063]	[0.067]	[0.048]	[0.049]	[0.054]
B. Logit Marginal Effects							
1. All Pairs (154 school-year averages)	0.418	0.142	0.154	0.176	0.064	0.068	0.085
		(0.063)	(0.062)	(0.055)	(0.053)	(0.053)	(0.049)
		[0.058]	[0.058]	[0.044]]	[0.050]	[0.050]	0.040
2. Balanced Pairs (122 school-year averages)	0.379	0.154	0.162	0.184	0.100	0.103	0.128
		(0.069)	(0.064)	(0.055)	(0.057)	(0.052)	(0.050)
		[0.062]	[0.058]	[0.044]]	[0.046]	[0.041]	[0.037]
3. Low-rate Pairs (110 school-year averages)	0.372	0.144	0.161	0.230	0.031	0.044	0.173
		(0.078)	(0.076)	(0.067)	(0.065)	(0.065)	(0.066)
		[0.070]	[0.069]	[0.050]	[0.057]	[0.060]	[0.058]

Notes: The table reports treatment effects estimated using school averages. The sample is limited to students in the upper half of the lagged score distribution using the classification from Table 4. All models stack average scores for 2001 and 2000 and control for school effects. The treatment effect is a treated-school dummy interacted with a 2001 year effect. Weighted estimates are weighted by school size. Conventional standard errors are reported in parentheses. Standard errors in brackets are robust (heteroscedasticity consistent).

Table 6: Mediating Outcomes

Outcome variable	Level or type	2001 control mean	Unweighted			Weighted		
			2000	2001	Stacked	2000	2001	Stacked
Units attempted (June 2001)	18	0.758	0.008	0.060	0.044	0.046	0.064	0.033
			(0.044)	(0.042)	(0.041)	(0.035)	(0.030)	(0.031)
	20	0.685	-0.005	0.100	0.098	0.032	0.087	0.070
			(0.045)	(0.045)	(0.044)	(0.038)	(0.034)	(0.034)
	22	0.613	0.005	0.105	0.093	0.045	0.083	0.054
			(0.050)	(0.051)	(0.050)	(0.045)	(0.037)	(0.039)
	24	0.509	0.025	0.103	0.073	0.046	0.061	0.028
			(0.052)	(0.055)	(0.047)	(0.048)	(0.041)	(0.038)
Units awarded (June 2002)	18	0.723	0.016	0.089	0.066	0.049	0.087	0.054
			(0.042)	(0.046)	(0.043)	(0.037)	(0.033)	(0.033)
	20	0.673	0.018	0.108	0.085	0.048	0.089	0.056
			(0.045)	(0.051)	(0.046)	(0.038)	(0.036)	(0.034)
	22	0.611	0.028	0.108	0.074	0.060	0.089	0.044
			(0.050)	(0.050)	(0.044)	(0.044)	(0.035)	(0.035)
	24	0.518	0.055	0.107	0.048	0.069	0.078	0.022
			(0.055)	(0.053)	(0.044)	(0.049)	(0.040)	(0.039)
Distribution requirement (June 2002)	Advanced subject	0.803	-0.005	0.046	0.046	0.012	0.024	0.031
			(0.039)	(0.050)	(0.039)	(0.034)	(0.033)	(0.028)
	Math	0.611	-0.070	0.048	0.117	0.009	0.029	0.037
			(0.054)	(0.052)	(0.057)	(0.047)	(0.040)	(0.043)
	English	0.692	0.083	0.099	0.007	0.136	0.078	-0.060
			(0.056)	(0.044)	(0.048)	(0.042)	(0.033)	(0.039)
	Writing	0.565	-0.032	0.023	0.053	-0.000	0.037	0.038
			(0.041)	(0.031)	(0.038)	(0.033)	(0.026)	(0.027)
Composite (June 2002)	20 awarded + math + adv. + writing	0.323	-0.051	0.093	0.146	0.016	0.064	0.070
			(0.050)	(0.051)	(0.054)	(0.044)	(0.039)	(0.037)

Notes: The table reports treatment effects on the outcomes listed in the first column, using models with school covariates and data for students in the upper half of the lagged score distribution. Estimates are from models estimated using data grouped by school, year, and score-group quartile (3^{rd} or 4^{th}), controlling for quartile main effects. Stacked estimates control for a full set of school effects instead of school covariates, and include a year dummy.

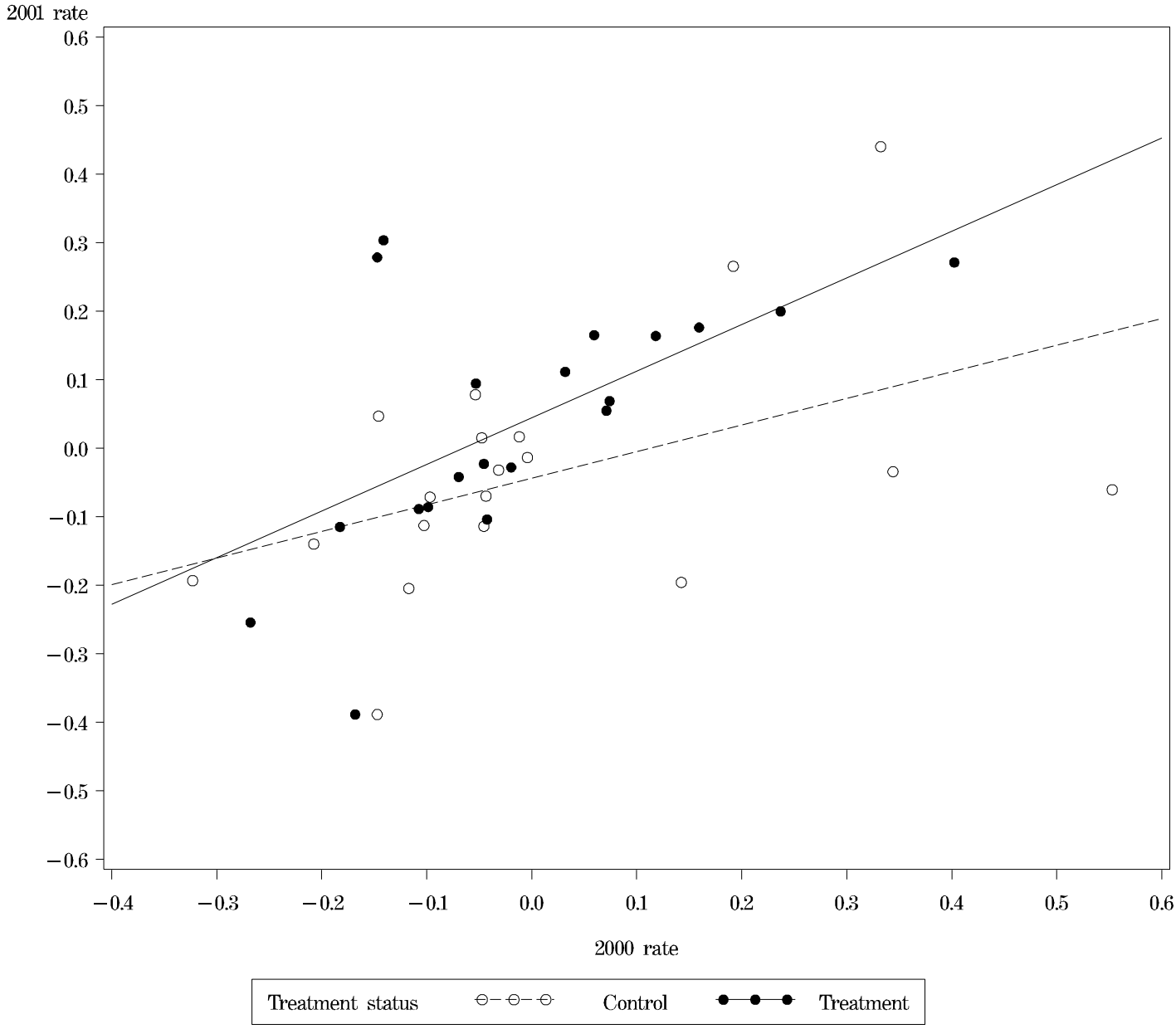

Figure 1. 2001 Bagrut rate vs. 2000 Bagrut rate by treatment status.
Residuals from regressions on school covariates.

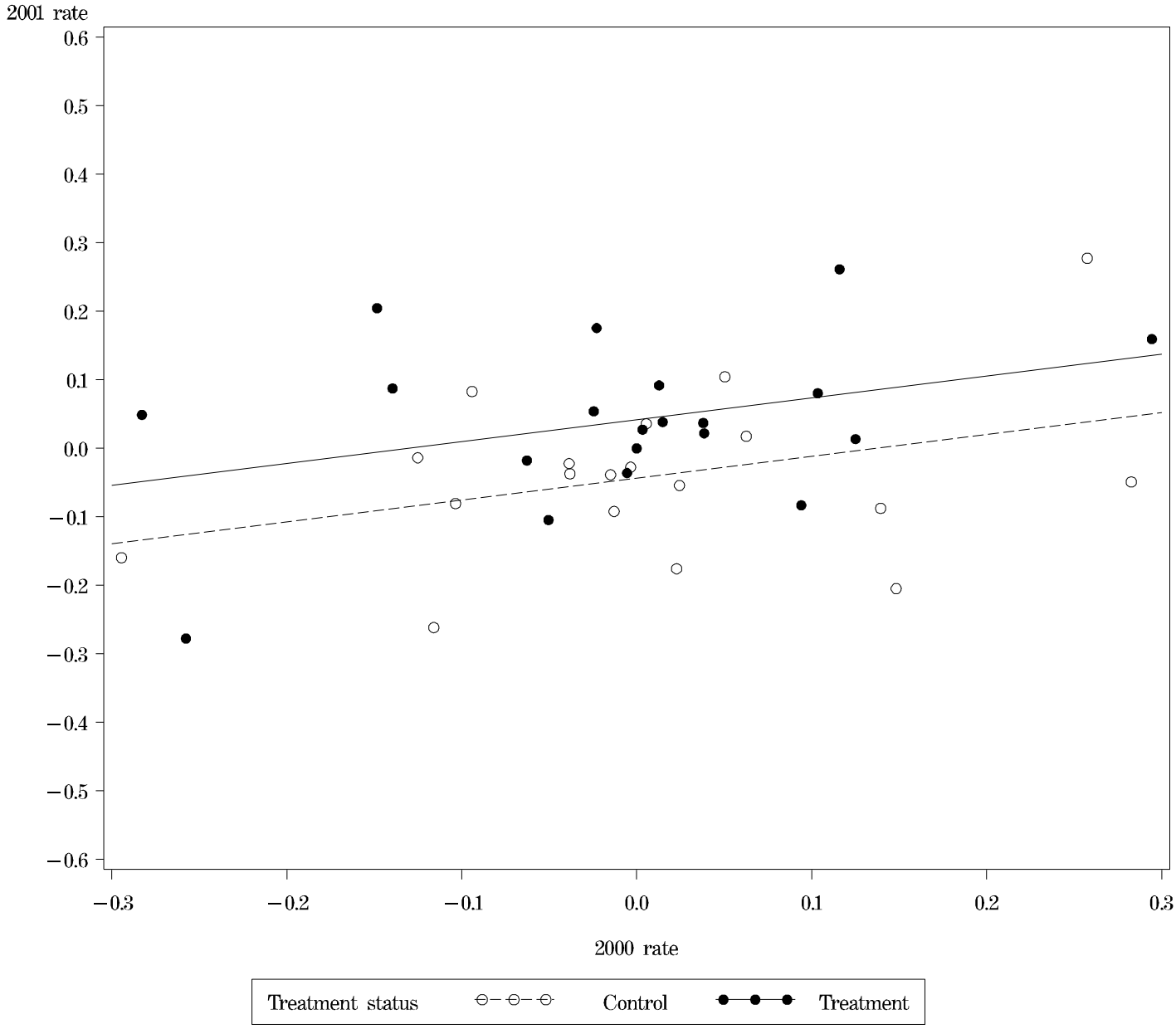

Figure 2. 2001 Bagrut rate vs. 2000 Bagrut rate by treatment status.
Residuals from regressions on school covariates and pair effects.