

C:\D_drive\projects\bonuses\revisions\AER_jan07\drafts\edited\AERrevision_sep07c.wpd; September 13, 2007

The Effects of High Stakes High School Achievement Awards:
Evidence from a Group-Randomized Trial

September 2007

The Effects of High Stakes High School Achievement Awards:
Evidence from a Group Randomized Trial

Abstract

In many countries, college-bound high school seniors must pass a test. In Israel, the national exit exam is known as the “Bagrut”, or matriculation certificate, obtained by passing a series of subject tests. The Bagrut is a pre-requisite for most post-secondary schooling. In spite of the Bagrut’s importance, 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. This intervention led to a substantial increase in certification rates for girls, though not for boys. The effects on girls were largely driven by an increase in Bagrut rates among those who had a relatively high *ex ante* chance of certification. The increase in Bagrut rates for this group translated in to an increased likelihood of college attendance five years later. Female Bagrut rates increased partly because treated girls were more likely to devote extra time to Bagrut preparation while boys essentially ignored the program.

I. Introduction

One of the most economically important educational milestones in many countries and in some American states is an exit exam for high school seniors, especially those intent on going to college. Examples include British A-levels, the French Baccalaureate, and the New York State Regents examinations. Since the 2001 No Child Left Behind Act, many US states have also adopted proficiency exams as a requirement for most high school graduates (see, e.g., Dee and Jacob, 2006, for a recent account). While American exit exams are not as closely linked to higher education as the European exams, public university systems in the US are increasingly likely to offer full tuition scholarships to high scorers.¹

In Israel, the high school matriculation certificate or Bagrut, awarded after a sequence of subject tests, is a formal pre-requisite for university admission and arguably marks the dividing line between the working class and the middle class. Our estimates from the 1995 Israeli Census suggest that even after controlling for highest grade completed, Bagrut holders earn 25 percent higher wages. Yet, in spite of the Bagrut's apparent economic and social value, Israeli society is marked by vast differences in Bagrut completion rates across regions and by socioeconomic background. As with high stakes exit exams in other countries, these disparities have led Israeli educators to implement remedial programs in an attempt to increase high school matriculation rates, with no apparent effect. These disappointing results echo similar findings from randomized trials in the U.S., where an array of service-oriented anti-dropout demonstrations for American teens have failed to increase high school 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 because of part-time work or other activities, 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 analyze the Achievement Awards demonstration, a project that provided cash awards for low-achieving high school students in Israel. The

¹For example, the University of Massachusetts offers free tuition to those with scores in the upper quartile of their district's MCAS score distribution.

intervention discussed here rewarded Bagrut completion and performance on Bagrut subject tests with direct payments to students.

Though the intervention studied here is unusual in some ways, there is growing interest in student incentive programs in primary and secondary education. Most visible among these is a nascent effort involving achievement incentives across the New York City school system, including a \$600 payment for each passing grade on New York State Regents exams (Medina, 2007). Our treatment also appears to be closely related to the Dallas Advanced Placement Incentive Program, which pays students (and teachers) for success on Advanced Placement exams. For a recent quasi-experimental evaluation of APIP see Jackson (2007).

Earlier demonstration projects in the same spirit include the Quantum Opportunities Program in US cities (Maxfield, Schirm, and Rodriguez-Planas, 2003); the Learning, Earning, and Parenting (LEAP) demonstration project in Ohio (Long, Gueron, Wood, Fisher, and Fellerath, 1996); Progreso 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 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

On balance, the findings reported here point to an increase in Bagrut rates in treated schools, the result of sharply increased certification rates for girls. The overall effect on girls is on the order of .10 (as compared

²As far as we know, ours and the study by Kremer, Miguel, and Thornton (2003) are the first completed projects using random assignment to evaluate substantial achievement-based payments to primary or secondary students using a randomized experimental design. Randomized trials investigating achievement incentives for college students include Angrist and Oreopoulos (2006), an ongoing experiment involving community college students (Bloom and Sommo, 2005), and Leuven, Oosterbeek, and van der Klaauw (2003).

³The National Merit programs give recognition and modest cash awards to a handful of high-achieving students based on their PSAT scores.

to a mean rate of about .29). Moreover, the increase for girls is driven by a group we think of as “marginal,” that is, girls with high predicted Bagrut rates *relative to other girls in our sample*. The prospects for this group are not rosy but they are not hopeless either. In particular, marginal students are those for whom certification is “within reach,” in the sense that adjustments in test-taking strategy or study time are likely to have a pay-off. We show, for example, that treated girls took only a few more tests but sharply increased their likelihood of meeting distribution requirements. Treated girls were also more likely to increase participation in a Spring study marathon. These findings points to more successful test-preparation and test-taking strategies.

The results naturally raise the question of whether the increased certification rate has a pay-off since additional study time and changes in test-taking strategy need not generate additional human capital. Importantly, therefore, we show that treated girls in the marginal group were substantially more likely to enroll in higher education five years after the intervention. As far as we know, ours is the first study to document this sort of long-run benefit in response to achievement awards.

The rest of the paper is organized as follows. Section II sketches some of the theoretical background motivating our intervention. Section III describes program details and discusses data and descriptive statistics. Section IV outlines the econometric framework. Section V discusses the effects on Bagrut rates and some of the mechanisms and channels for these effects, including the differential response by gender. This analysis suggests that female Bagrut rates increased partly because treated girls were more likely to devote extra time to Bagrut preparation, while boys essentially ignored the program. Section VII reports the results on post-secondary schooling outcomes and Section VI concludes.

II. Theoretical Context

Why do young men and women fail to complete high school? Why don't more go to college? These questions present something of a puzzle since the economic returns to schooling appear to be very high, and seem likely to 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). Liquidity constraints also surface in the literature on market work and study-time during school (e.g., Tyler, 2003; Stinebrickner and Stinebrickner, 2003). This literature also notes that the causal links between student time allocation and educational outcomes are hard to quantify using observational data, though most of the evidence suggests student effort matters to at least some extent (Hotz, *et al*, 2002; Stinebrickner and Stinebrickner, 2004, 2007).

A number of features of the Israeli economic and social environment dovetail with the issues raised in previous research on low educational attainment. First, many Israeli students work, especially in poorer areas.⁴ Work may come 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.

The experimental program discussed here, which we refer to as the Achievement Awards demonstration, was motivated by a desire to tip the scales towards current investment in schooling and away from market work, vocational training, or leisure. This 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 and the recently-launched Opportunity NYC initiative, which offers students cash awards for achievement on New York State

⁴Roughly 36 percent of surveyed boys and 23 percent of surveyed girls in our sample reported working for pay in the last 6 months of the school year. Boys worked 31 hours per month and girls worked 13 hours per month (including zeros for non-workers).

Regents exams. Keane and Wolpin (2000) simulated the impact of the Reich policy in the context of a structural model of education choice. They estimated that this program would have a large impact on high school graduation rates and college attendance, especially for blacks.

Bagrut status can be understood in the American vernacular as a “college-bound” indicator. Most of the Israeli students who fail to complete a Bagrut still finish their secondary schooling. Nevertheless, post-secondary schooling options for high school graduates without a Bagrut are limited; very few will obtain further schooling. Even institutions that are not otherwise very selective, such as teachers’ colleges and two-year professional programs for nursing, optometry, and computer programming, favor applicants with a Bagrut. Consistent with this, regression evidence from the Israeli Census suggests that the economic returns to a Bagrut are high, though we do not have a good experiment for the earnings consequences of Bagrut certification. Clearly, however, if certification increases schooling it very likely increases earnings. A recent quasi-experimental study of exit exams in Texas suggests that those who pass these exams go on to get more post-secondary schooling than they otherwise would have (Martorell, 2005).

The Bagrut is awarded on the basis of a challenging series of exam - arguably harder than many US exit exams - and many Israeli students may not be able to complete a Bagrut no matter how hard they try. The Bagrut requires students to pass a number of tests (mostly in 12th grade) and to satisfy distribution requirements showing proficiency in Reading (Hebrew), Math, English, and an advanced subject of their choosing. An important feature of our investigation, however, is a focus on students we see as *on the margin for success*, that is, they are close to Bagrut certification (from the point of view of policy-makers).⁵ Among students from the control schools in our sample, about 43 percent finish school with the minimum credit-units criterion for Bagrut status (20 units) satisfied, yet only 48 percent of these end up with a Bagrut, either because they are a couple of units away or because of a failure to meet distribution requirements. For the purposes of our study, we define marginal students *ex ante* by predicting Bagrut success in the control group using exams taken in 10th and 11th

⁵This 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 nationally.

grade. The average probability of success in the marginal group is close to 37 percent for boys and 52 percent for girls, in contrast with only a 3-6 percent success rates for non-marginal students. The substantial cross-sectional and time-series variation in Bagrut rates suggests that under some circumstances, students in this marginal group can clear the final Bagrut hurdle.

III. Program Details

A. The Israeli School System

High school students earn a Bagrut 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. Thus, Bagrut certificates are typically obtained at the end of senior year or later. 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. A minimum of 20 credit units is required to qualify for a matriculation certificate, though some study programs require more, and students must also satisfy the distribution requirements. 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 ended up receiving a Bagrut certificate. In our sample, however, Bagrut rates are much lower.

B. Research Design and Program Implementation

In December 2000, Ministry of Education officials (in consultation with us) selected the 40 non-vocational high schools with the lowest 1999 Bagrut rates in a national ranking, but above a minimum threshold rate of 3 percent. Some low-rate schools in the relevant universe were ineligible to participate in the experiment for technical or administrative reasons such as poor organization or data. We also felt that schools with virtually no Bagrut recipients were unlikely to benefit from the program. The list of participating schools

included 10 Arab and 10 Jewish religious schools.⁶ The total number of treated schools was determined by the program budget constraint. Ultimately, about \$650,000 (3.1 million shekels) was awarded.

Treatment was randomly assigned to 20 of the 40 participating schools. While not large enough to ensure treatment-control balance, the number of clusters used here is typical of other GRTs (see, e.g., Feng, *et al* [2001] or Donner, Brown, and Brasher [1990]). To improve treatment-control balance the assignment 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. Bagrut rates from 1999 were used to select and match schools within pairs because the 2000 data were incomplete when treatment was assigned. Treatment was assigned randomly within pairs, the most common matching strategy in GRTs (see, e.g., Gail, *et al*, 1996). Thus, the treatment design did not balance Bagrut rates in the immediate pre-treatment year, though balancing was near perfect in 1999. The raw treatment-control difference in the 1999 Bagrut rates is about -.018.

Every student in treated schools who received a Bagrut was eligible for a payment. Randomized trials that assign treatment status to entire schools are often 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, successful implementation of educational interventions depends partly on the cooperation of teachers and school administrators, and the intervention may get additional leverage from peer effects when classmates participate.

The timing of the program and key data collection points are summarized in a chart in the appendix. The orientation for principals was in January 2001, about one-third of the way into the 2000-2001 school year. Follow-up contacts in March 2001 were used to verify the participation by contacting principals and school administrators. Five treatment schools are non-compliers in the sense that, following the orientation session

⁶Israel runs separate school systems for Secular Jews, Religious Jews, and Arabs. Rules and standards for Bagrut certification are similar in all three.

in January 2001, principals in these schools had taken no concrete actions to inform pupils or teachers about the program and/or indicated that they did not wish to participate.

Program Parameters

The program was meant to last 3 years, with awards given to high school students in every grade. Seniors became aware of the program about one-third into the year, before the “big push” Bagrut study effort that is traditional in the Spring. Modest 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 taking any Bagrut component test, regardless of the outcome, with NIS1,500 to be given for passing component tests before senior year. The planned award schedule also appears in the appendix. 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 dropout, and about twice as much as a student might earn working full-time in two summer months.⁷

In practice, implementation of the program as originally conceived was affected by a number of events. First, following an election and a change in government, a new minister of education was appointed. Second, a budgetary crisis in late Fall of 2001, the beginning of the second year of the program, led to a sharp reduction in the education budget. Because of these events and perhaps also due to adverse publicity when the program became public knowledge, the award scheme for younger cohorts was eventually canceled. As a consequence, awards were given for only one year of achievement and the maximum amount awarded was NIS6,000. Although 10th and 11th graders were eligible for more modest short-term awards, by the time the bulk of their Bagrut effort took place, the program had ceased. This disruption notwithstanding, the program for high school seniors operated as planned, and the post-intervention survey shows that most students were aware of specific

⁷ These estimates are based on the minimum wage rate for the 17-18 age group in 2001 obtained taken from a company that compiles these data over time [see <http://www.hilan.co.il/calc/MinimumWageCalculator.aspx>].

program features . The analysis in this paper is therefore limited to high school seniors.

C. Data and Descriptive Statistics

Baseline data were collected in January 2001 while the main Bagrut outcome for the treated cohort comes from tests taken in June of 2001. A small follow-up survey was conducted in late summer and early Fall of 2001. Students had an unanticipated opportunity to be retested in August-September of 2001 and a regularly scheduled second chance in the winter of 2002. Though the results using winter 2002 data are similar, we prefer the June data because of the disruption and uncertainty introduced by the Bagrut retests which were unexpectedly offered in the Fall of 2001.

In addition to Bagrut outcomes, our administrative data set includes basic socioeconomic information. This information is summarized in Table 1, which presents 2000-2001 means for our sample of schools and the nation. By construction, the Bagrut rate in the experimental sample is much lower; 22-24 percent versus 61-63 percent nationally (among schools with a positive Bagrut rate in 1999). Relative to the national average, the experimental sample includes more students attending Arab schools, but fewer attending religious schools. Students in the experimental sample also have less educated parents, more siblings, and are more likely to be new immigrants than in the country as a whole.⁸

Appendix Table A1 reports Bagrut means and enrollment counts for 1999-2002 in each of the 39 schools that participated in the study (the control school in pair 6 had closed by the time treatment was assigned). The 1999 Bagrut rates used for matching ranged from 3.6-28.6 percent and are (by design) similar within pairs. Schools ranged 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. For example, many absorb large cohorts of new immigrants.⁹

⁸About 10-15 percent of the administrative records are missing socioeconomic characteristics. We imputed missing data using means by sex and school type. However, data on our core outcome variables, Bagrut status and post-secondary schooling, are essentially complete.

⁹The variability in Bagrut rates in later years 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, although it still worth something. The R^2

IV. Econometric Framework

The following model is used to estimate treatment effects from Bagrut data for individual students:

$$y_{ij} = \Lambda[X_j' \alpha + \sum_q d_{qi} \delta_q + W_i' \beta + \gamma z_j] + \epsilon_{ij}, \quad (1)$$

where i indexes students and j indexes schools, and $\Lambda[\cdot]$ is a possibly nonlinear link function (in this case, the logistic transformation or the identity). We assume that $E[y_{ij}] = \Lambda[X_j' \alpha + \sum_q d_{qi} \delta_q + W_i' \beta + \gamma z_j]$, where the expectation is conditional on individual and school characteristics (alternately, this is the minimum mean square error approximation to the relevant conditional expectation). School-level variables include the treatment dummy, z_j , and a vector of school-level controls, X_j , that includes a dummy for Arab schools and a dummy for Jewish religious schools. In some specifications, this vector also includes dummies for randomization pairs. The vector W_i , contains individual characteristics such as parental schooling, the number of siblings, and immigrant status. Some models also include a function of lagged test scores which, as we show below, predicts Bagrut status exceptionally well. This function consists of three dummies $\{d_{qi}; q=2, 3, 4\}$ indicating the quartile of a student's credit-unit-weighted average test score on tests taken before January 2001, when the program was implemented.¹⁰ We also estimated models replacing score-group dummies with a linear term in lagged scores.

The first econometric issue raised by our study is non-compliance. In follow-up contacts in March 2001, we verified the level of compliance by contacting the 39 participating principals and school administrators (one school had closed). 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. Because the decision to cooperate may be related to potential outcomes, even within pairs, we analyze the data based on randomly-assigned intention to treat, i.e., the

from a weighted regression of the 2001 rate on the 1999 rate is .15. It bears emphasizing that substantial variability in year-to-year performance measures for individual schools is not unique to our sample. 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.

¹⁰In particular, we calculated each student's credit-unit-weighted average score as of January (coding zeros for those with no tests) and then divided students into quartiles on the basis of this weighted average. Students were assigned to quartiles using the distribution of credit-unit weighted average scores for their cohort in our sample.

reduced-form impact of the randomized offer of program participation, indicated by z_j in equation (1). The conclusion briefly discusses the impact of adjustments for non-compliance. Because the compliance rate was high, this involves a modest re-scaling of the reduced-form estimates.

A second statistical issue is how best to adjust inferential procedures for clustering at the school level, a consequence of the GRT research design. 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 parametric cluster adjustments overly optimistic (Thornquist and Anderson, 1992; Feng, *et al*, 2001).

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. The advantage of GEE are flexibility and availability in proprietary software. 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. As with parametric Moulton-type or design-effect adjustments, GEE standard errors are also biased downwards (See, e.g., Wooldridge, 2003).

To sidestep the problem of downward-biased GEE standard errors, we estimated standard errors (for models without school effects) using Bell and McCaffrey’s (2002) Biased Reduced Linearization (BRL) estimator. 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. Appendix Table A2 compares standard errors for linear probability models similar

to equation (1), estimated using alternative cluster adjustments. This table show that the BRL standard errors, while slightly larger than those produced by the conventional GEE cluster adjustment, are similar to those arising from a two-step procedure based on adjusted group means proposed by Donald and Lang (2007). Since an analysis of data grouped at the cluster level is likely to be conservative, this similarity is encouraging. Table A2 also serves as a robustness check in the sense that the basic findings are indeed apparent in a school-level analysis of group means.

A final statistical issue worth discussing is the role of pair effects, which are sometimes included in the vector of school-level controls, x_j . In principle, pair effects can be dropped without biasing the estimates of treatment effects since intention to treat is a (fair) coin toss in each pair. Moreover, ignoring stratification variables may lead to more precise estimates since the inclusion of pair effects uses up degrees of freedom (Diehr, *et al*, 1995; Angrist and Hahn, 2004). Therefore, because the pair effects in our design turn out to explain little of the variation in the dependent variable, estimates from models with and without pair effects are discussed below.

V. Results

A. Cross-section estimates

Estimates of equation (1) support the notion that the Achievement Awards program increased Bagrut rates in 2001, though there is also some evidence of spurious effects in outcomes for the previous (2000) cohort. The increase in Bagrut rates is due entirely to an increase for girls. These findings can be seen in Table 2, which reports OLS and Logit estimates of equation (1) using the full sample, as well as separate results for boys and girls.¹¹ Panel A reports results for 2001, the post-treatment year, while Panels B and C reports results for the 2000 and 2002 cohorts as a specification check. The first set of results shown in each panel is from models that include school covariates; the second is from models that adds lagged score quartile dummies and

¹¹The logit results are reported as marginal effects on the treated. The sample used for Logit drops schools with zero dependent variable means.

individual student characteristics (mother and father's schooling, the number of siblings, and student's immigrant status). Each of these two specifications was estimated with and without pair effects as additional controls.

Estimates from all models for the combined sample of boys and girls, reported in columns 1-2, are positive, though only those from models with a full set of controls and pair effects are (marginally) significant. For example, the OLS estimates in column 1 (no controls) is .056 (s.e.=.049). The estimate with a full set of school controls, lagged score quartiles, and individual controls is larger and more precise, at .067 (s.e.=.036). The corresponding set of Logit marginal effects, reported in column 2, are slightly smaller. There is some evidence of imperfect randomization, however, since the estimates for 2000 are also positive. The largest of the estimates for 2000, in the first row of panel B, is almost as large as the corresponding estimate for 2001. On the other hand, the gap between the 2001 and 2000 estimates increases when additional control variables are included in both models, and none of the estimates for 2000 are significantly different from zero.

Separate analyses by sex show sharp differences in effects for boys and girls. The estimates for boys, reported in columns 3-4, are uniformly small and negative; none are significantly different from zero, while the estimates for boys in 2000 are small, positive, and also insignificant.¹² For example, 2001 estimates from models with all controls are -.022 (OLS) and -.023 (Logit). In contrast, the 2001 estimates for girls are on the order of .10, and most are at least marginally significantly different from zero. Moreover, while the 2000 estimates for girls are also positive, none of these are as large as those for 2001, and all are insignificant.

It's also worth noting that the basic pattern of 2001 and 2000 results in Table 2 is apparent in an analysis of school-level grouped means, reported in Appendix Table A2. For example, the 2001 weighted grouped-data estimate for girls from a model with school covariates (adjusted for lagged test scores using the Donald and Lang [2007] two-step procedure) is .095 (s.e.=.044). Grouped-data estimates for girls without covariates and without weighting are similar.

¹²Pair effects are omitted from models estimated separately for boys and girls so as not to lose pairs that include single-sex (religious) schools.

A causal interpretation of the 2001 effects for girls is further reinforced by the analysis of data from the 2002 graduating cohort. This can be seen in Panel C, which shows that Bagrut rates were remarkably similar in treatment and control schools in the year after the experiment ended.¹³ The estimated treatment-control differences are on the order of $-.02$, for both boys and girls. Although seniors in the 2002 cohort were offered small payment to take and pass at least one Bagrut subject test as 11th graders in 2001, no further incentives were offered to this cohort since the program was cancelled before they began their senior year. Moreover, we found no evidence that the modest payments offered to 11th graders affected their test-taking behavior or results. Thus, the treatment experienced by the 2002 cohort can be seen as providing a sort of placebo control in that these students attended treated schools, but were exposed to little in the way of a changed environment.

The estimates in Table 2 suggest the Achievement Awards program had no effect on boys, but show reasonably clear evidence of increased Bagrut rates for girls. The picture is muddled somewhat by the positive (though insignificant) effects on girls in 2000, though the 2002 results suggest that the imbalance in 2000 may have been transitory. Nevertheless, in an effort to reduce any possible bias from school-level omitted variables, we estimated models using stacked 2000, 2001, and 2002 data in a set-up that includes school effects. This procedure controls for any time-invariant school-level omitted variable that might explain higher Bagrut rates in 2000 in treated schools. The introduction of school (fixed) effects also provide an alternative approach to the clustering problem. Before turning to the models with school fixed effects, however, we refine the estimation strategy by isolating the group of marginal students most likely to benefit from the Achievement Awards intervention. We first report results for marginal groups using models for levels and then turn to models for marginal groups incorporating school effects.

¹³The 2002 Bagrut levels are higher than those for 2000-2001 because the 2002 data reflect additional attempts at certification after the main testing round for high school seniors. This difference should not affect the comparison of treatment and control groups, however.

B. Identification of Marginal Students

Groups of marginal students were defined using a predictive regression that models the probability of success as a function of school characteristics and individual covariates. The predictive model is:

$$y_{ij} = \Lambda[X_j' \pi_0 + \sum_q d_{qi} \pi_q + W_i' \pi_1] + \eta_{ij}, \quad (2)$$

where X_j , W_i , and $\{d_{qi}; q=2, 3, 4\}$ are as in equation (1). The model is meant solely as a classification device, we estimated it using the 2001 control sample only. Some of the specifications omit mother's schooling from W_i since the parents' schooling effects are never both significantly different from zero.

The logit coefficients estimated using equation (2), reported in Table 3, show that the three lagged score quartile dummies are far and away the best predictor of Bagrut status. Especially noteworthy is the fact that the lagged score coefficients dwarf family background effects, both in size and statistical significance.¹⁴ This is not surprising, since the marginal probability of Bagrut certification in June 2001 was about 1% in the lowest score quartile, 9% in the second score quartile, 29% in the third score quartile, and 49% in the upper quartile. This gradient reflects the fact that Bagrut status is determined in large part by accumulating credit units. Students entering senior year with very few units simply cannot make up the shortfall. On the other hand, the odds of Bagrut success are substantial for students who've done well on 10th and 11th grade subject tests. Conditional on lagged scores, SES is of modest value as a predictor of Bagrut status, though background covariates do more for girls than for boys.

Motivated by these results, we re-estimated equation (1) using two subgroup classification schemes. The first splits students according to the lagged score distribution, again using credit-unit-weighted scores. In other words, we divided students into two roughly equal-sized groups, the top half with $d_{3i}=d_{4i}=1$. Second, we used the fitted values from model (2), again dividing students into roughly equal-sized groups. This second scheme provides a check on the notion that high lagged scores identify students who have a shot at certification. For both schemes, we estimated models that include school covariates and either a quartile main effect calculated from the distribution of the classification variable, or a linear term in the classification variable.¹⁵

¹⁴Here we show logit coefficients instead of marginal effects since it is the *relative* predictive power of covariates that is of primary interest. Students were divided into lagged-score or predicted-bagrut groups based on models or score distributions within gender and year.

¹⁵The results are almost identical when a scaled score is substituted for the credit-unit lagged raw score used to define classification groups.

Both classification schemes appear to do a good job of isolating students likely to be affected by treatment, a fact documented in the descriptive statistics at the top of Table 4. Both schemes divide the sample into a low-achieving group that has less than a 5 percent chance of certification and a relative high-achieving group. Girls in the top group have a better than 50 percent probability of Bagrut success.

Panel A of Table 4, which reports treatment effects in 2001, consistently shows small and insignificant estimates for both boys and girls in the bottom half of the distribution of lagged scores or fitted values from equation (2). In contrast, the estimates for girls in the top half of the distribution, reported in odd-numbered columns, show large significant effects. For example, the estimated effect on girls in the top half of the distribution of lagged scores is .206 (s.e.=.079), while the corresponding effect on girls in the top half of the distribution of fitted values is .194 (s.e.=.077). Moreover, the corresponding estimates for girls in 2000, reported in Panel B, are less than half as large and none is significantly different from zero. Especially encouraging is the fact that the estimates using 2002 data, reported in Panel C, are essentially zero for both boys and girls, in both the top and bottom half of the lagged-score or predicted-success distribution.¹⁶

C. Models with School Effects

Table 4 presents a clear pattern of significant treatment effects for girls in the upper half of the 2001 lagged score distribution, with no significant effects in other years. On the other hand, the presence of some fairly large treatment-control differences in 2000 raises a concern about omitted school effects. We therefore estimated stacked models controlling for additive school fixed effects. The coefficient of interest in the stacked specification is the interaction between a dummy for 2001, and the treatment indicator, z_j . The resulting estimates can be interpreted as a student-weighted difference-in-differences procedure comparing treatment effects across years (except that the number of schools differ from year to year).¹⁷ Models with school effects

¹⁶We also explored models allowing interactions with individual characteristics, with inconclusive results. We were not able to find a subgroup of boys exhibiting a pattern of strong treatment effects similar to that for girls.

¹⁷The differences-in-differences analogy is imperfect for two reasons. First, the estimates are implicitly weighted by the number of students in each school. Second the panel is unbalanced across years because a few schools that were open in one year were closed in another. In addition, any school with a mean Bagrut rate of zero must be dropped from the Logit estimation.

control for time-invariant omitted variables, a particular concern given the positive estimates in 2000 data. Moreover, school effects provide an alternative control for school-level clustering and absorb some of the variability in average Bagrut rates by school, possibly leading to a gain in precision.¹⁸

The stacked equation used to estimate models with school effects can be written:

$$y_{ijt} = \Lambda[\mu_j + \xi_t + X_j' \alpha_t + W_i' \beta + \gamma(z_j d_t)] + \epsilon_{ijt}, \quad (3)$$

where $t=2000-2002$; $d_t=1[t=2001]$; μ_j is a school and ξ_t is a year effect, and α_t is a year-specific vector of coefficients on school covariates. The micro covariates, W_i , now include either a dummy for first quartile students (in the bottom half) or third quartile students (in the top half) or a linear term in lagged score or predicted Bagrut success, depending on how the sample was divided into higher and lower achievers. The dependent variable, y_{ijt} , is the Bagrut status of student i in school j in year t .

Not surprisingly given the baseline differences in 2000, estimates of the model with school effects using 2001 and 2000 data are smaller than the corresponding estimates for 2001 only. This can be seen in Panel A of Table 5. The estimated matriculation gains for girls in the upper half of the lagged score distribution are about .093 (s.e.=.043) in a model with a dummy for third-quartile students (column 3) and .102 (s.e.=.043) in a model with linear control for the lagged score. The corresponding estimates when students are split by predicted Bagrut rates (column 7) are .08 and .09, respectively. The estimated effects on girls in the bottom half of the distribution of the classification variable is small and insignificantly different from zero. The estimates for boys are close to zero in all subgroups.

Panel B of Table 5 reports results from a stacked specification that uses 2002 data as a control instead of 2000. Because the 2002 treatment-and control Bagrut rates are almost perfectly balanced, these results show somewhat larger estimates than the 2001-2000 stack. For example, the estimates for girls in the top half of the lagged score distribution climb to about 17 percent, while the estimates for boys in both halves, as well as for girls in the bottom-half sample, are again zero. Finally, estimation using 2000-2002 data produces results

¹⁸BRL standard errors adjust *inference* for the uncertainty that arises with omitted random effects, while models with school fixed effects change the estimator.

between those from the 2000-01 and 2001-02 stacked samples, with a modest gain in precision. For example, the estimates for girls in the top half of the lagged score distribution are .13-.14, with a standard error of .039.

A final refinement on this investigation of models with school effects looks at estimates in the 3rd and 4th quartiles of the lagged score or predicted Bagrut distribution. These results, reported in Table 6, tell a story similar to that in Table 5, except that the largest effects for girls in the stacked models now appear in the upper *quartile* of the classification distribution. For example, estimated in stacked models without linear control for lagged scores, the effect on girls in the upper (fourth) quartile of the lagged score distribution is .145 (s.e.=.06), while the corresponding estimate in the third quartile is .028 (s.e.=.064). This is evidence for the notion that the group most likely to benefit from short-term incentives are those for whom the target is within reach. (This is apparently a necessary but not sufficient condition, since boys in the upper quartile were unaffected by incentives).

Although the effects on girls in the upper lagged score quartile reported in Table 6 are substantial, these results translate into a modest overall impact after averaging across all treated students (i.e., including both boys and girls, and all lagged score quartiles). Table 6 therefore suggests we have succeeded in using previous test performance to zero in on a subgroup where a fair number of students can be nudged into passing the Bagrut with a little extra effort or more focused test-taking behavior. This is consistent with Israeli Ministry of Education reports showing that a substantial number of students - over 20 percent nationally - “almost” get a Bagrut in the sense of fulfilling most but not all of the Bagrut requirements. For example, many students are only a few units short while others have the requisite number of units but fail to meet distribution requirements. The next subsection explores the anatomy of Bagrut success for treated students in greater detail.

D. Channels for Improvement

To obtain a Bagrut, students must clear a number of hurdles. These include subject tests worth a minimum of 20 or more credit units, a Writing Composition requirement, and Math and English requirements. We therefore looked at what might be considered the proximate causes of Bagrut success: whether students

were tested for more units, were more likely to succeed on the exams they took, and whether they were more likely to satisfy distribution requirements. The first outcome in this context is the number of 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.

Estimates of effects on the number of credit units attempted show a small increase for treated girls with lagged scores in the upper half of the 2001 score distribution. This can be seen in Table 7, which reports estimated treatment effects on indicators for units-attempted thresholds for boys and girls in the top half of the lagged score distribution. The estimates were constructed using models similar to those used to construct the estimates in Tables 3 and 4, but with different dependent variables. In particular, we look at effects on indicators for at least 18, 20, 22, or 24 units attempted (with attempts measured as of June 2001). The results for girls show a pattern of positive though small and mostly insignificant effects on attempted units. The largest effect, on attempts of 20 or more units, is .073 (s.e.=.038). Estimated effects on boys' attempts are smaller and none are close to significance.¹⁹

Paralleling the increase in units attempted around the 20-unit margin is a small increase in the number of units awarded to girls, as can be seen in the next panel down in Table 7.²⁰ For example, the estimated effect on the probability of obtaining 18-plus units is .053 (s.e.=.035) while the effect on the likelihood of obtaining 20-plus units is .064 (s.e.=.038). Although they suggest a modest increase in units awarded, these effects are too small to explain the program treatment effect on Bagrut rates. It seems likely, therefore, that the net program effect comes partly through other channels.

The means in Table 7 show that 69 percent of the upper-half subsample of girls in 2001 obtained 22 credit units. Many therefore fail to get certified because the units are in the wrong subjects. Consistent with

¹⁹The estimates for the 2000-2001 levels and stack in Table 7 use the same sample so they can be more easily compared. The marginal effects in Table are evaluated at the average success rate for treated schools in both years, also in the interests of comparability.

²⁰These are estimates on indicators for having been awarded 18, 20, 22, and 24 units measured as of June 2002. The timing here has to do with data quality issues. For details, see Angrist and Lavy (2004).

this, part of the increase in Bagrut rates appears to have arisen through an increased likelihood of satisfying distribution requirements in Math and English. These affects are documented in the Table 7 panel labeled “distribution requirements”. For example, the Achievement Awards program led to a .081 (s.e.=.041) increase in the likelihood girls met the Math requirement and a .11 (s.e.=.039) increase in the likelihood girls met the Writing requirement.

The estimated program effects on distribution requirements are stronger and more clearly indicative of an increase in student effort than the effects on units attempted and awarded. The impact on Math and Writing requirements is most likely the result of a shift in effort towards these specific subject areas. The program impact is also reflected in an increase in the likelihood of Bagrut success conditional on units attempted. The conditional results, reported at the bottom of Table 7, show that treated girls who were tested for at least 18 units were .098 (s.e.=.047) more likely to obtain a Bagrut, with somewhat smaller effects in other conditioning groups.²¹ The effects on distribution requirements and on Bagrut success conditional on attempts suggests that the Achievement Awards program elicited more than a mechanical response involving additional test-taking alone.

Mechanisms and Gender Differences

Why do girls respond to incentives while boys do not? We investigated three explanations that seem relevant in our context. The first is that even within classification groups, girls may be more likely to be “marginal” in the sense of being close to the threshold for a Bagrut. We evaluated this possibility by looking at a number of definitions of near-Bagrut categories in the 2001 control data and in the 2000 data (for example, students with 20 units and all but one of the distribution requirements satisfied). As it turns out, boys and girls are about equally likely to fall into groups that are relatively close to certification.

A second possibility is program awareness. Data from the Ministry of Education follow-up survey,

²¹It should be noted, however, that the conditional-on-attempts results do not have a simple causal interpretation where there is also an impact on attempts. Angrist, Bettinger, and Kremer (2005) document the phenomenon of relatively weak marginal test-takers in a quasi-experimental study of the effects of school vouchers on college admissions tests.

conducted late summer and early fall of 2001, can be used to investigate this possibility. The survey data are far from perfect but the results are suggestive. We find that girls are more likely to report having been aware of the program (*ex post*), especially among students in the top half of the lagged score distribution. In this sample, 61 percent of girls and 54 percent boys demonstrated program awareness. This difference does not seem large enough to explain the boy-girl differences in treatment effects, however. Also, program awareness expressed *ex post* might simply reflect the higher award rates for girls (since those who got an award are more likely to recall being in the program.)

Finally, and perhaps most relevant, the survey data includes measures of study time, study effort and hours worked in paid employment. There is little evidence of a difference in these variables between treatment and control groups. However, a common practice in Israeli high schools is for students and teachers to get together in marathon study sessions around the holidays (Hanukkah in Winter and Passover in Spring). The Passover marathon is dedicated to a big push for the Bagrut. In our sample, girls are more likely to participate in these marathon study sessions than boys, especially in the top half of the lagged score distribution (30 percent for girls vs. 19 percent for boys). Among upper-quartile girls (the group with the largest treatment effect for Bagrut rates), we also find significantly higher Passover marathon participation among treated girls (an effect of .193, s.e.=.085) in this sample. Importantly, there is no treatment effect on participation in the Hanukkah marathon (since this predates treatment).

The findings on extra study time and distribution requirements suggest that some girls responded to incentives with a focused and well-timed study effort while boys did not. On the other hand, these results do not provide a deep psychological explanation of female responsiveness. It is worth noting, however, that there is a literature suggesting that adolescent girls have more self-discipline (e.g., Duckworth and Seligman, 2006) and are more likely to delay gratification (e.g., Silverman, 2003) than adolescent boys. Among young adults, Warner and Pleeter (2001) find that male enlisted personnel behave as if they have higher discount rates than women in the same group.

There is also a consistent pattern of stronger female response to education incentives, with the evidence

coming from a surprising variety of countries and settings. Especially relevant are recent studies of tuition aid by Dynarski (2005) and tuition penalties by Garibaldi, *et al* (2006), both of which find larger effects for females. Also closely related is a recent randomized trial looking at cash payments for academic achievement among college freshman; this study finds clear effects for females but no effect on males (Angrist and Oreopoulos, 2006). A more modest but still marked gender differential crops up in the response to randomly assigned vouchers for private secondary schools in Colombia (Angrist, *et al*, 2002). These vouchers incorporated an incentive component because voucher retention was conditional on academic performance.²²

VI. Post Secondary Schooling

This section discusses estimates of the effects of the Achievement Awards program on enrollment in post-secondary schooling. The post-secondary academic schooling system in Israel includes seven universities (one of which confers only graduate and PhD degrees), over 40 colleges that confer academic undergraduate degrees (some of these also give Masters degrees), and dozens of teachers' colleges that confer Bachelor of Education or practical engineering degrees.²³ The national enrollment rates for the cohort of graduating seniors in 1995 (through 2001) was 52.7 percent, of which 20.2 (38) percent were enrolled in universities, 14.8 (28) percent were enrolled in general colleges and 14.7 (28) percent were enrolled in teachers' and practical engineering colleges.²⁴

The outcome variables of interest are indicators of ever having enrolled in post-secondary institutions of various types as of the 2006-7 school year. We measure this outcome for our 2000 and 2001

²²Somewhat farther afield, Anderson (2005) shows that three well-known early childhood interventions (Abecedarian, Perry, and the Early Training Project) had substantial short and long term effects on girls but no effect on boys. Similarly, a number of public-sector training programs generated larger effects on women than men (Lalonde, 1995). The recent Moving to Opportunity housing study likewise generated clear benefits for girls, with little or even adverse effects on boys (Kling, Leibman, and Katz, 2005).

²³Practical engineering colleges run 2-3 year programs awarding degrees or certificates in fields like electronics, computers, and industrial production. Two additional years of schooling in an engineering school are required in order to complete a B.Sc. degree in engineering. A 1991 reform sharply increased the supply of post secondary schooling in Israel by creating publicly funded regional and professional colleges. New institutions granting academic degrees are supervised by the council for Higher Education that supervises the seven research universities.

²⁴ These data are from the Israel Central Bureau of Statistics, *Report on Post Secondary Schooling of High School Graduates in 1989-1995*, [available at: http://www.cbs.gov.il/publications/h_education02/h_education_h.htm].

high school cohorts. Because of compulsory military service, many of the students from these cohorts who enrolled in post secondary schooling will not have graduated by the 2006-7 academic year.²⁵ We therefore focus on enrollment instead of completion.

Our information on enrollment in post secondary schooling comes from administrative records provided by Israel's National Insurance Institute (NII). The NII is responsible for social security and mandatory health insurance in Israel. The NII tracks post-secondary enrollment because students pay a lower health insurance tax rate. Post-secondary schools are therefore required to send a list of enrolled students to the NII every year. For the purposes of our project, the NII Research and Planning Division constructed an extract containing the 2001-2007 enrollment status of students in our study. This file was merged with the other information in our sample and used for analysis at the NII headquarters in Jerusalem.

We coded three indicators for enrollment in post secondary schooling. The first identifies enrollment in one of the seven universities (at any time from 2001-7); the second expands this definition to include certified academic colleges; the third adds teachers' and practical engineering colleges to the second group. All universities and colleges require a Bagrut for enrollment. Most teacher's and practical engineering colleges also require a Bagrut, though some look at specific Bagrut components without requiring full certification.

To avoid an over-abundance of results given the number of post-secondary outcomes of interest, we focus on results from our preferred specification. These results, reported in Table 8, are from models similar to equation (3), i.e., estimated using stacked 2000 and 2001 data including school fixed effects. Stacked estimates are reported for the top and bottom half of the lagged score distribution as in columns 1-4 of Table 5 (estimates conditional on probability of success are omitted). We also report results from a further split by quartile, as in Table 6. Our discussion also focuses on Logit marginal effects, with the exception of effects on university enrollment. Because there are many schools with no students attending

²⁵Boys serve for three years and girls for two (longer if they take a commission). Ultra-orthodox Jews are exempt from service as long as they are enrolled in seminary (Yeshiva), orthodox Jewish girls are exempt upon request, and Arabs are exempt though some volunteer.

university, the results for university enrollment are from linear models. For comparability, we report both OLS and Logit estimates for other outcomes.

Consistent with the fact that few of the students in our study end up in one of Israel's research universities, there is no effect of treatment on university enrollment. This can be seen in Panel A of Table 8. University enrollment rates are low even in the top half of the lagged score distribution, 0.064 for boys and .069 girls. Enrollment rates in the bottom half of the lagged-score distribution are essentially zero for boys and 1.1 percent for girls. The treatment effects for both genders and in both halves of the distribution are also zero. Although enrollment rates are higher for students in the upper quartile of the lagged score distribution, around 10-11 percent for girls, there is still no program effect on this outcome (as can be seen in columns 5-8).

Combined college and university enrollment rates are much higher than university enrollment rates alone, as shown in Panel B of Table 8 (labeled "All Academic"). Panel B also shows some evidence of a program effect on college enrollment for girls in the top quartile of the lagged score distribution. For example, the program effect on college or university enrollment of girls in the 4th quartile is .086 (s.e.=.055), which can be compared to a mean of .248. Moreover, widening the definition of post-secondary study to include teachers colleges and practical engineering colleges leads to larger and broader effects, reported in Panel C. The Logit estimate of the effect on the enrollment of girls using the most inclusive enrollment is .081 in the top half of the lagged score distribution (s.e=.038; mean=.331) and .128 (s.e=.061) in the top quartile (where the mean is .429). These estimates are slightly higher in models with linear control for lagged scores. In contrast, the estimates for boys are mostly small and insignificant for all enrollment outcomes in all subsamples.

The effects on the most inclusive enrollment outcome in the upper half and upper quartile of the lagged score distribution are about three-fourths the size of the effects on Bagrut rates when the latter are estimated using the same model and sample. For purposes of this comparison, Bagrut results for the

relevant sample appear at the bottom of Table 8. These estimates are slightly smaller than those reported in Table 5 (for top half girls) and slightly larger than those reported in Table 6 for upper quartile girls).

The Bagrut results in Table 8 suggest that three-quarters of the additional Bagrut rates received as a consequence of the Achievement Awards intervention caused additional post-secondary enrollment of some type (e.g., compare .12 in Panel C of column 8 with .16 in Panel D). Although this proportion may seem high, it bears emphasizing that the overall post-secondary enrollment rate for all Israeli Bagrut holders is also high, on the order of 78 percent for the 1994/95 cohort.

A related interpretative point is the fact that non-compliance does not affect the ratio of post-secondary effects to Bagrut effects. This ratio can be interpreted as an instrumental variables estimate of the effect of Bagrut certification on post-secondary enrollment. The instrumental variables adjustment for non-compliance implicitly divides both post-secondary and Bagrut results by the same take-up rate. On the other hand, the program impact on certification is not the only channel by which the Achievement Awards program may have increased post-secondary enrollment. In other words, Bagrut certification need not satisfy an exclusion restriction for the reduced-form program effect on post-secondary outcomes. Some students may have had better post-secondary options by virtue of increasing the number of units tested or by satisfying distribution requirements even if their Bagrut status was unaffected.

VII. Summary and Conclusions

A randomly-assigned offer of cash awards to students in low-achieving schools appears to have generated substantial gains in the matriculation rates of girls. Although there is some imbalance in the Bagrut rates from the year preceding treatment, a causal interpretation of the results is supported by estimates from models that control for unobserved school effects, and by the absence of a treatment effect in the cohort that graduated after the one treated. The overall impact on girls is driven by treatment effects in a group we see as marginal; that is, students relatively close to certification thresholds. The effect on this group (girls in the upper half of the lagged score distribution) are around .10 in a model allowing for

omitted school effects. This is our best guess of the program impact on the marginal group of girls. There appears to be no effect on boys or on girls who are not in the marginal group, so that the overall program effect is small.

The program effects become somewhat larger when allowance is made for the fact that one-quarter of schools offered the opportunity to participate in the program either declined to participate or failed to provide rosters in time. Adjusting for non-compliance, the effect on treated girls (i.e., girls in treated schools) is about one-third larger (1/.75) than the reduced-form intention-to-treat effects discussed in the paper. On the other hand, the intention-to-treat effect may be a better gauge of future program impact since it seems likely that other programs of this sort will allow for non-compliance and give school administrators the opportunity to opt-out.

An analysis of the channels through which students may have responded to incentives generates some evidence of increased effort in the form of more exams taken or more difficult exams attempted and especially an increased likelihood of meeting distribution requirements. This turns up in a higher success rate for girls conditional on the number of exams attempted. Using survey data, we also find girls increased their exam study time in the pre-Bagrut holiday period. Boys did not respond in any way that we can detect. The gender differential in program response echoes male-female differences in the response to financial incentives for college achievement and in the response to tuition subsidies and penalties in a number of recent demonstrations.

We have also shown that for many students, the increase in matriculation rates translated into increased post-secondary enrollment using our broadest measure of post-secondary studies. The sharpest boost is for girls in the top quartile of the lagged score distribution, a 12 percentage point increase, while the post-secondary enrollment gain for girls in the top half of the lagged score distribution is about 7 percentage points. These increases seem likely to have generated substantial economic gains since the returns to post-secondary education in Israel appear to be high. For example, Frisch and Moalem (1999)

estimate the return to a year of college to be about 11 percent in the late 1990s, while Frisch (2007) estimates the average return to having any post secondary schooling to be about 34 percent.

Because of the substantial economic return to post-secondary education, the incentive scheme used here is likely to generate a net social gain. The bonus offer of NIS6,000 shekels for matriculating seniors 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 roughly \$385 per treated student. Using the average annual earnings of those with 11-12 years of schooling in 2005 as a base, (\$14,910), and assuming that it takes 10 years before any benefits are realized (to allow for military service, college attendance, and labor-market entry), the internal rate of return for investment in Achievement Awards is about 8.6 percent.²⁶ This suggests that cash incentives of this sort can make economic sense even without taking distributional implications into account. More focused programs, e.g. programs targeting girls and/or marginal students, could well generate even higher economic returns.

²⁶ This calculation is based on the following assumptions: The estimates in Table 8 suggest the program raised college attendance by say .10 in the top half of the girls lagged score distribution (taking the smaller of the two for the broadest category). This implies an average enrollment effect (including boys) of 0.025. Assuming the schooling generated by this enrollment boost amounts to two years, each yielding an 11 percent rate of return, the program effect is worth $.025 * .11 * 2 = .0055$ percent. Based on an average annual earnings of \$14,910, the gain per year is \$82. The earnings data are from Table 22 in [http://www.cbs.gov.il/publications/income_05/pdf/t22.pdf].

Table 1: Descriptive Statistics

	Experimental Sample			National		
	All (1)	Boys (2)	Girls (3)	All (4)	Boys (5)	Girls (6)
	A. 2001					
Bagrut Rate	.243	.200	.287	.629	.574	.678
<i>School Covariates</i>						
Arab School	.348	.374	.320	.163	.159	.167
Religious School	.115	.084	.148	.170	.154	.184
<i>Micro Covariates</i>						
Father's education	10.1 (3.07)	9.82 (3.11)	10.3 (3.00)	12.2 (3.48)	12.2 (3.48)	12.1 (3.48)
Mother's education	10.0 (3.29)	9.87 (3.32)	10.2 (3.24)	12.0 (3.42)	12.0 (3.42)	11.9 (3.42)
Number of Siblings	3.74 (2.66)	3.65 (2.64)	3.84 (2.68)	2.97 (1.95)	2.91 (1.91)	3.03 (1.98)
Immigrant	.064	.029	.100	.023	.021	.025
Lagged Score	53.1 (29.4)	52.1 (29.4)	54.2 (29.3)	---	---	---
<i>Proportion of Missing Values</i>						
Father's education	.144	.168	.118	.124	.128	.120
Mother's education	.153	.173	.132	.136	.142	.130
Number of Siblings	.116	.111	.122	.107	.110	.105
Observations	3,821	1,960	1,861	76,990	36,423	40,567
	B. 2000					
Bagrut Rate	.224	.177	.272	.611	.560	.657
<i>School Covariates</i>						
Arab School	.319	.352	.286	.161	.160	.163
Religious School	.134	.098	.170	.171	.154	.186
<i>Micro Covariates</i>						
Father's education	9.87 (3.07)	9.75 (3.15)	10.0 (2.99)	12.1 (3.56)	12.1 (3.57)	12.0 (3.56)
Mother's education	9.80 (3.26)	9.71 (3.33)	9.9 (3.18)	11.9 (3.48)	11.9 (3.50)	11.9 (3.45)
Number of Siblings	3.68 (2.47)	3.53 (2.34)	3.84 (2.58)	2.99 (1.98)	2.92 (1.92)	3.06 (2.03)
Immigrant	.074	.039	.109	.032	.029	.035
Lagged Score	50.2 (28.9)	49.1 (29.4)	51.4 (28.4)	---	---	---
<i>Proportion of Missing Values</i>						
Father's education	.109	.121	.096	.087	.094	.080
Mother's education	.115	.129	.100	.085	.094	.077
Number of Siblings	.101	.105	.098	.103	.107	.100
Observations	4,039	2,038	2,001	77,241	36,484	40,757

Notes: Columns 1-3 report sample means. Standard deviations are shown in parentheses. Statistics in columns 4-6 are from the authors' tabulation of administrative data, for schools with a positive Bagrut rate in 1999.

Table 2 - Estimated 2001 Treatment Effects and Specification Checks

	Pair effects	Boys + Girls		Boys		Girls	
		OLS (1)	Logit (2)	OLS (3)	Logit (4)	OLS (5)	Logit (6)
A. 2001							
<i>Dep. Mean</i>		.243		.200		.287	
Sch covs	No	.056 (.049)	.051 (.045)	-.010 (.052)	-.011 (.055)	.105 (.061)	.093 (.053)
	Yes	.052 (.047)	.054 (.043)	---	---	---	---
Sch covs + quartiles + micro covs	No	.052 (.039)	.047 (.039)	-.022 (.043)	-.023 (.045)	.105 (.047)	.097 (.046)
	Yes	.067 (.036)	.055 (.036)	---	---	---	---
<i>Number of students</i>		3,821		1,960		1,861	
<i>Number of schools</i>		39		34		34	
B. 2000							
<i>Dep. Mean</i>		.224		.177		.272	
Sch covs	No	.050 (.056)	.046 (.051)	.045 (.060)	.040 (.055)	.075 (.067)	.069 (.061)
	Yes	.043 (.059)	.045 (.058)	---	---	---	---
Sch covs + quartiles + micro covs	No	.030 (.041)	.018 (.042)	.009 (.050)	.006 (.052)	.066 (.046)	.051 (.046)
	Yes	.043 (.044)	.030 (.046)	---	---	---	---
<i>Number of students</i>		4,039		2,038		2,001	
<i>Number of schools</i>		39		33		35	
C. 2002							
<i>Dep. Mean</i>		.305		.257		.357	
Sch covs	No	-.019 (.071)	-.019 (.071)	-.026 (.073)	-.028 (.075)	-.010 (.077)	-.010 (.078)
	Yes	-.018 (.050)	-.018 (.059)	---	---	---	---
Sch covs + quartiles + micro covs	No	-.023 (.044)	-.021 (.045)	-.026 (.046)	-.024 (.047)	-.015 (.046)	-.014 (.046)
	Yes	-.027 (.033)	-.033 (.034)	---	---	---	---
<i>Number of students</i>		4,328		2,269		2,059	
<i>Number of schools</i>		38		33		33	

Notes: The table reports OLS estimates and Logit marginal effects. Panel A shows treatment effects. Results from 2000 and 2002 are specification checks. BRL standard errors are reported in parentheses.

Table 3 - Determinants of Bagrut Status, 2001

	Boys+Girls		Boys		Girls	
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Dep. Mean</i>	.219		.194		.239	
<i>School Covariates</i>						
Arab school	.987 (.540)	1.08 (.558)	.625 (.830)	.901 (.905)	1.30 (.420)	1.26 (.400)
Religious school	.627 (.579)	.632 (.578)	2.56 (.603)	2.57 (.628)	.282 (.508)	.281 (.508)
<i>Micro Covariates</i>						
Father's education	.076 (.044)	.056 (.040)	.081 (.053)	.016 (.047)	.068 (.039)	.074 (.038)
Mother's education	---	.034 (.037)	---	.109 (.079)	---	-.012 (.032)
Has more than 4 siblings	-.431 (.307)	-.422 (.308)	.080 (.521)	.093 (.547)	-.682 (.254)	-.685 (.255)
Immigrant	.924 (.383)	.905 (.385)	1.31 (1.204)	1.35 (1.196)	1.10 (.364)	1.12 (.364)
<i>Lagged Score Quartiles</i>						
2nd	2.83 (.739)	2.84 (.738)	1.77 (.871)	1.78 (.880)	3.16 (.784)	3.16 (.785)
3rd	4.34 (.662)	4.35 (.664)	3.81 (.788)	3.86 (.776)	4.69 (.781)	4.69 (.781)
4th	5.02 (.581)	5.01 (.583)	4.57 (.629)	4.56 (.625)	5.33 (.738)	5.33 (.735)
Observations	1,876		850		1,026	
Number of schools	19		15		18	

Notes : The table reports logit estimates. The estimates in this table were constructed using the sample of control schools only. BRL standard errors are reported in parentheses.

Table 4 - Levels Estimates in Covariate Subgroups

	By Lagged Score				By Predicted Probability			
	Boys		Girls		Boys		Girls	
	Top (1)	Bottom (2)	Top (3)	Bottom (4)	Top (5)	Bottom (6)	Top (7)	Bottom (8)
A. 2001								
<i>Dep. Mean</i>	.365	.035	.518	.056	.368	.032	.518	.056
<i>Models</i>								
Sch covs + group main effect	-.013 (.083)	.007 (.016)	.206 (.079)	-.020 (.024)	-.047 (.077)	.005 (.016)	.194 (.077)	-.015 (.023)
Sch covs + group linear control	-.009 (.083)	.007 (.017)	.213 (.079)	-.021 (.022)	-.044 (.079)	.001 (.017)	.207 (.078)	-.019 (.026)
<i>Number of students</i>	980	980	933	928	980	980	932	929
B. 2000								
<i>Dep. Mean</i>	.318	.035	.475	.068	.320	.033	.478	.066
<i>Models</i>								
Sch covs + group main effect	.055 (.079)	-.014 (.035)	.098 (.074)	.009 (.027)	.033 (.078)	.004 (.027)	.086 (.071)	.009 (.023)
Sch covs + group linear control	.055 (.079)	-.014 (.035)	.094 (.072)	.007 (.026)	.010 (.077)	.000 (.028)	.089 (.070)	.007 (.024)
<i>Number of students</i>	1,022	1,016	1,004	997	1,021	1,017	1,002	999
C. 2002								
<i>Dep. Mean</i>	.475	.040	.611	.101	.472	.042	.608	.106
<i>Models</i>								
Sch covs + group main effect	-.018 (.101)	-.004 (.016)	-.017 (.088)	-.030 (.032)	-.029 (.098)	-.007 (.017)	-.006 (.078)	-.021 (.029)
Sch covs + group linear control	-.008 (.097)	-.003 (.016)	-.013 (.088)	-.037 (.031)	-.032 (.088)	-.015 (.021)	-.001 (.073)	-.020 (.028)
<i>Number of students</i>	1,135	1,134	1,035	1,024	1,135	1,134	1,030	1,029

Notes: The table reports Logit marginal effects in top and bottom subgroups, classified by lagged test scores or predicted probability of Bagrut success. Panel A shows treatment effects. Results from 2000 and 2002 are specification checks. BRL standard errors are reported in parentheses.

Table 5 - Stacked Estimates in Covariate Subgroups, with School Fixed Effects

	By lagged score				By Predicted Probability			
	Boys		Girls		Boys		Girls	
	Top (1)	Bottom (2)	Top (3)	Bottom (4)	Top (5)	Bottom (6)	Top (7)	Bottom (8)
A. Stacked 2000 and 2001								
<i>Dep. Mean</i>	.341	.050	.497	.074	.344	.048	.498	.072
<i>Models</i>								
Sch covs + group main effect	-.043 [.045]	-.035 [.039]	.093 [.043]	-.065 [.035]	-.030 [.045]	-.069 [.042]	.082 [.043]	-.046 [.035]
Sch covs + group linear control	-.035 [.046]	-.031 [.038]	.102 [.043]	-.052 [.031]	-.006 [.043]	-.077 [.044]	.091 [.043]	-.050 [.035]
<i>Number of students</i>	2,002	1,395	1,931	1,613	2,001	1,355	1,930	1,639
B. Stacked 2001 and 2002								
<i>Dep. Mean</i>	.424	.052	.569	.087	.424	.051	.568	.090
<i>Models</i>								
Sch covs + group main effect	-.010 [.041]	.017 [.018]	.165 [.045]	-.007 [.019]	-.017 [.042]	.013 [.015]	.144 [.046]	-.008 [.020]
Sch covs + group linear control	-.001 [.041]	.016 [.019]	.168 [.045]	-.006 [.019]	-.013 [.041]	.014 [.014]	.159 [.046]	-.011 [.021]
<i>Number of students</i>	2,115	1,532	1,958	1,778	2,115	1,541	1,952	1,782
C. Stacked 2000, 2001 and 2002								
<i>Dep. Mean</i>	.390	.047	.539	.082	.390	.046	.539	.080
<i>Models</i>								
Sch covs + group main effect	-.022 [.038]	.009 [.018]	.133 [.039]	-.019 [.020]	-.019 [.039]	.000 [.017]	.118 [.039]	-.015 [.020]
Sch covs + group linear control	-.012 [.038]	.009 [.018]	.139 [.039]	-.016 [.019]	-.007 [.037]	.000 [.017]	.129 [.039]	-.019 [.020]
<i>Number of students</i>	3,137	2,463	2,948	2,692	3,136	2,471	2,941	2,815

Notes: The table reports Logit marginal effects estimated in models with school fixed effects. The treatment effect is an interaction between a dummy for treated schools and year=2001. Estimates are for subsamples classified as in Table 4. Robust standard errors are shown in brackets.

Table 6 - Levels and Stacked Estimates
By Lagged Score and Predicted Probability Quartiles

	Lagged Score Quartiles				Predicted Probability Quartiles			
	Boys		Girls		Boys		Girls	
	4th (1)	3rd (2)	4th (3)	3rd (4)	4th (5)	3rd (6)	4th (7)	3rd (8)
A. 2001								
<i>Dep. Mean</i>	.454	.272	.616	.420	.500	.227	.648	.384
<i>Models</i>								
Sch covs	-.025 [.043]	.005 [.041]	.291 [.052]	.143 [.044]	-.089 [.044]	.005 [.038]	.221 [.051]	.172 [.042]
Sch covs + group linear control	-.031 [.043]	.005 [.041]	.299 [.052]	.136 [.044]	-.091 [.044]	.006 [.038]	.232 [.051]	.177 [.041]
<i>Number of students</i>	502	478	466	467	508	472	474	458
B. 2000								
<i>Dep. Mean</i>	.433	.203	.618	.331	.427	.214	.630	.325
<i>Models</i>								
Sch covs	.078 [.044]	.040 [.033]	.117 [.047]	.088 [.039]	.039 [.044]	.034 [.035]	.089 [.047]	.097 [.038]
Sch covs group linear control	.077 [.044]	.040 [.034]	.114 [.047]	.079 [.039]	.004 [.044]	.019 [.036]	.087 [.045]	.099 [.038]
<i>Number of students</i>	510	512	503	501	511	510	503	499
C. Stacked 2000 and 2001								
<i>Dep. Mean</i>	.443	.257	.618	.370	.468	.219	.641	.357
<i>Models</i>								
Sch covs	-.029 [.061]	-.102 [.078]	.145 [.060]	.028 [.064]	-.029 [.065]	-.060 [.062]	.151 [.063]	.041 [.073]
Sch covs + group linear control	-.029 [.061]	-.103 [.079]	.152 [.061]	.038 [.062]	-.001 [.064]	-.033 [.058]	.163 [.063]	.047 [.071]
<i>Number of students</i>	1,010	912	965	922	1,009	981	971	940

Notes: The table reports Logit marginal effects in upper-quartile and third-quartile subgroups, classified by lagged test scores or predicted probability of Bagrut success. Robust standard errors are shown in brackets.

Table 7 - Mediating Outcomes

Outcome variable	Boys				Girls			
	2001	2000	2001	Stacked	2001	2000	2001	Stacked
	Mean				Mean			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Units attempted</i>								
18	.749	.059	.049	.012	.849	.064	.091	.024
		(.055)	(.065)	[.038]		(.056)	(.044)	[.033]
20	.700	.050	.063	.035	.793	.034	.127	.073
		(.063)	(.066)	[.040]		(.054)	(.053)	[.038]
22	.630	.066	.052	.002	.717	.042	.106	.050
		(.073)	(.063)	[.041]		(.066)	(.054)	[.042]
24	.536	.073	.042	-.011	.578	.033	.065	.020
		(.075)	(.069)	[.042]		(.081)	(.054)	[.043]
<i>Units awarded</i>								
18	.728	.061	.057	.014	.804	.097	.156	.053
		(.055)	(.068)	[.039]		(.059)	(.053)	[.035]
20	.686	.064	.059	.016	.762	.077	.150	.064
		(.061)	(.067)	[.041]		(.057)	(.059)	[.038]
22	.622	.052	.046	.009	.688	.108	.150	.049
		(.072)	(.062)	[.043]		(.065)	(.058)	[.041]
24	.527	.094	.046	-.036	.590	.071	.118	.045
		(.080)	(.065)	[.044]		(.079)	(.066)	[.043]
<i>Distribution requirements</i>								
Math	.557	-.007	.004	.020	.685	.062	.153	.081
		(.082)	(.063)	[.044]		(.074)	(.059)	[.041]
English	.707	.107	.082	-.001	.771	.143	.111	-.009
		(.062)	(.057)	[.040]		(.071)	(.048)	[.035]
Writing	.700	.014	-.003	.002	.815	-.002	.117	.106
		(.058)	(.062)	[.040]		(.055)	(.050)	[.039]
<i>Bagrut, Conditional on Units Attempted</i>								
18	.488	.047	-.044	-.085	.605	.095	.200	.098
		(.090)	(.087)	[.053]		(.086)	(.093)	[.047]
20	.519	.050	-.064	-.121	.641	.128	.191	.063
		(.095)	(.088)	[.055]		(.092)	(.100)	[.046]
22	.556	.032	-.055	-.083	.664	.127	.199	.054
		(.106)	(.100)	[.058]		(.097)	(.104)	[.046]
24	.589	.008	-.055	-.065	.711	.051	.176	.086
		(.117)	(.102)	[.061]		(.096)	(.100)	[.052]

Notes: The table reports Logit marginal effects estimated in models with school covariates and a lagged score main effect, using data for students in the upper half of the lagged score distribution. The marginal effects for the stacked data are based on mean predicted values for treated students in 2000 and 2001. BRL standard errors are reported in parentheses (for levels estimates). Robust standard errors are reported in brackets (for stacked estimates).

Table 8 - Stacked Estimates of the Effects on Education in the 2000 and 2001 Cohorts
By Covariates Subgroups

	By Lagged Score Halves				By Lagged Score Quartiles			
	Boys		Girls		Boys		Girls	
	Top (1)	Bottom (2)	Top (3)	Bottom (4)	4th (5)	3rd (6)	4th (7)	3rd (8)
A. University								
<i>Dep. Mean</i>	.064	.003	.069	.011	.101	.025	.112	.027
<i>Models</i>								
OLS, Sch covs (t.v) + group linear control	.019 [.022]	.004 [.004]	-.021 [.023]	.014 [.008]	.041 [.038]	.014 [.022]	.012 [.042]	-.034 [.020]
B. All Academic								
<i>Dep. Mean</i>	.171	.041	.176	.046	.230	.110	.248	.107
<i>Models</i>								
OLS, Sch covs (t.v) + group linear control	.009 [.033]	.001 [.018]	.028 [.033]	-.006 [.020]	.002 [.051]	.035 [.040]	.086 [.055]	-.031 [.039]
Logit, Sch covs (t.v) + group linear control	.022 [.026]	.005 [.019]	.045 [.033]	.018 [.019]	.009 [.046]	.045 [.030]	.086 [.055]	-.006 {.065}
C. Academic, Teachers and Practical Engineering								
<i>Dep. Mean</i>	.252	.081	.331	.099	.304	.198	.429	.236
<i>Models</i>								
OLS, Sch covs (t.v) + group linear control	-.041 [.038]	-.018 [.025]	.067 [.040]	.035 [.028]	-.042 [.055]	-.042 [.052]	.123 [.060]	.031 [.052]
Logit, Sch covs (t.v) + group linear control	-.028 [.035]	-.012 [.028]	.081 [.038]	.036 [.019]	-.032 [.051]	-.031 [.056]	.128 [.061]	.047 [.051]
D. Bagrut (replication over the National Insurance non-missing)								
<i>Dep. Mean</i>	.341	.034	.490	.070	.443	.236	.611	.372
<i>Models</i>								
OLS, Sch covs (t.v) + group linear control	-.027 [.039]	-.011 [.016]	.102 [.042]	-.049 [.024]	-.022 [.060]	-.063 [.051]	.161 [.060]	.038 [.059]
Logit, Sch covs (t.v) + group linear control	-.038 [.046]	-.038 {.040}	.108 {.043}	-.053 [.032]	-.029 [.061]	-.116 [.080]	.163 {.062}	.037 {.060}
<i>Number of students</i>	1,997	1,985	1,882	1,711	1,007	988	921	958

Notes: The table reports OLS coefficients and Logit marginal effects for stacked models with school fixed effects. Columns 1-4 show estimates in top and bottom subgroups as in Table 5. Columns 5-8 show estimates in upper and third-quartile subgroups as in Table 6. Robust standard errors are reported in brackets. Conventional standard errors are reported in braces where robust Logit standard error estimation failed. Where both are available, the conventional and robust standard errors are virtually identical. The means and number of observations reported relate to the OLS estimates.

APPENDIX: ACHIEVEMENT AWARDS PROGRAM RULES AND TIMING

Program Rules

1. Award schedule

<u>Grade</u>	<u>Milestone</u>	<u>Reward (NIS)</u>
10	Tested for at least 1 unit and enrolled in 11th grade	500
	Passed this test	1500
11	Tested for at least 3 units and 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
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

Schools Selected and Principals informed	December	2000
Orientation for principals and students	January	2001
Baseline administrative data collected	January	2001
Media Coverage	May	2001
Bagrut Tests	June	2001
Student Survey	August/September/October	2001
Re-test (Math and English)	August-September	2001
Winter Retest	December-January	2001- 2002

Notes: in March 2001 principals were interviewed to determine whether the program was publicized in schools. Bonuses were paid in May 2002.

Table A1 - Descriptive Statistics for the School Experiment

Pair	Treated	Non complier	Arab school	Relig. School	Enrollment				Bagrut Passing Rate			
					1999	2000	2001	2002	1999	2000	2001	2002
1			X		153	173	175	249	.046	.000	.091	.185
1	X			X	56	59	45	43	.036	.051	.000	.047
2				X	242	170	147	88	.054	.094	.184	.034
2	X				179	185	145	158	.050	.108	.110	.095
3					88	99	71	80	.114	.000	.056	.075
3	X		X		123	129	99	103	.098	.054	.030	.068
4					81	68	73	67	.148	.162	.082	.075
4	X		X		187	223	248	297	.134	.390	.339	.458
5					125	124	96	70	.152	.105	.083	.129
5	X			X	55	39	38	48	.145	.077	.579	.167
6	X				117	125	123	154	.171	.136	.154	.273
7				X	16	28	16	22	.188	.214	.375	.545
7	X			X	67	85	58	63	.179	.165	.483	.444
8				X	57	48	61	60	.193	.771	.328	.583
8	X				90	97	113	106	.189	.186	.168	.368
9					61	40	59	60	.197	.350	.000	.383
9	X			X	10	14	9	21	.200	.071	.667	.429
10	X	X		X	34	39	26	43	.206	.410	.654	.488
10	X				135	135	108	102	.207	.267	.361	.441
11	X				136	148	134	169	.213	.176	.164	.172
11	X				129	158	152	159	.209	.165	.092	.151
12			X		19	24	20	60	.211	.667	.250	.617
12	X			X	32	44	24	20	.219	.250	.500	.350
13					146	118	119	137	.219	.153	.185	.219
13	X				85	80	86	114	.224	.363	.372	.342
14					208	170	185	199	.236	.153	.276	.352
14	X	X	X		75	50	64	120	.227	.560	.484	.367
15			X		156	153	163	202	.244	.176	.331	.391
15	X	X			138	141	152	171	.254	.610	.467	.520
16			X		102	115	108	107	.255	.226	.213	.327
16	X				74	61	75	74	.257	.098	.107	.095
17				X	23	14	16	0	.261	.071	.000	---
17	X		X		76	68	67	62	.263	.441	.448	.435
18			X		216	211	219	246	.273	.303	.301	.305
18	X	X			200	148	110	146	.275	.162	.173	.103
19					141	111	77	183	.284	.541	.636	.776
19	X	X			123	40	62	43	.276	.025	.081	.093
20					185	161	111	94	.286	.161	.126	.181
20	X		X		144	144	167	188	.285	.389	.353	.309

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. The treatment school in pair 6 is meant to be grouped with pair 7. The 1999 data was recorded at the time of matching.

Table A2: Estimates Using Micro Data for the Schools Experiment

	Two-Step Procedure							
	Un-weighted			Weighted			Micro Data	
	No	Sch	Sch Cov	No	Sch	Sch Cov	Sch	Sch Cov
	Controls	Cov	+ Pair	Controls	Cov	+ Pair	Cov	+ Pair
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
	A. 2001							
1. All	.102 (.053) [.051]	.106 (.048) [.046]	.111 (.051) [.034]	.045 (.040) [.038]	.055 (.038) [.036]	.064 (.040) [.025]	.055 (.036) {.039}	.064 (.025) {.036}
2. Boys	.054 (.060) [.058]	.036 (.050) [.047]	---	.015 (.048) [.044]	-.003 (.042) [.042]	---	-.004 (.042) {.046}	---
3. Girls	.110 (.049) [.047]	.114 (.048) [.047]	---	.090 (.046) [.048]	.095 (.045) [.044]	---	.095 (.043) {.047}	---
	B. 2000							
1. All	-.009 (.050) [.049]	-.007 (.049) [.047]	-.003 (.051) [.033]	.028 (.043) [.043]	.032 (.043) [.041]	.040 (.046) [.031]	.031 (.041) {.043}	.040 (.030) {.045}
2. Boys	-.008 (.053) [.055]	.000 (.052) [.054]	---	.024 (.045) [.046]	.031 (.046) [.048]	---	.031 (.048) {.052}	---
3. Girls	.042 (.056) [.054]	.047 (.056) [.052]	---	.056 (.051) [.050]	.059 (.051) [.045]	---	.058 (.044) {.048}	---

Notes: Columns 1-6 report estimates constructed using the Donald and Lang (2007) two-step procedure, where the first step adjusts school fixed effects (means) for micro covariates, and the second step is a group-level regression using the adjusted means. The micro covariates in the first step are dummies for lagged score quartiles. Conventional standard errors for the second step are shown in parentheses. Heteroscedasticity-consistent standard errors for the second step are reported in brackets. Columns 7 and 8 report regression results using micro data, with controls for lagged score quartiles. The standard errors in parenthesis in columns 7-8 are adjusted for school clustering using formulas in Liang and Zeger (1986) similar to Stata's "cluster" command. Standard errors in braces in columns 7-8 use MacCaffrey and Bell's (2002) BRL estimator.

Table A3 - Parameter Estimates for the Stacked Regression in Table 5 (Linear Control, Top Half Sample, by Lagged Score)

	Stacked 2000 and 2001		Stacked 2001 and 2002	
	Boys	Girls	Boys	Girls
<i>Dep. Mean</i>	.341	.497	.424	.569
Treated	-.173 [.226]	.527 [.220]	-.006 [.220]	.840 [.225]
Arab School x (Year=2001)	-.175 [.227]	.254 [.236]	-.198 [.222]	.377 [.230]
Religious School x (Year=2001)	1.76 [.376]	-.354 [.355]	1.22 [.365]	1.17 [.371]
Lagged Score	.048 [.006]	.073 [.007]	.062 [.006]	.062 [.007]
(Year=2001)	.118 [.206]	-.409 [.195]	-.519 [.200]	-.958 [.195]
<i>Number of students</i>	2,002	1,931	2,115	1,958

Notes: The table reports Logit coefficients for the model indicated. Robust standard errors are shown in brackets.

REFERENCES

- Anderson, Michael, "Uncovering Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," MIT Department of Economics, mimeo, September 2005.
- 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, Eric Bettinger, and Michael Kremer, "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *American Economic Review*, 2006,
- 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, Joshua, Daniel Lang, and Philip Oreopoulos, "Lead Them to Water and Pay Them to Drink: An Experiment with Services and Incentives for College Achievement," NBER Working Paper 12790, December 2006.
- Angrist, Joshua D. and Victor Lavy, "The Effect of High Stakes High School Achievement Awards: Evidence from a School-Centered Randomized Trial," IZA Working Paper 1146, May 2004.
- 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).
- Bloom, Dan and Colleen Sommo, "Building Learning Communities: Early Results from the Opening Doors Demonstration at Kingsborough Community College," New York: MDRC, 2005.
- Central Bureau of Statistics, *Statistical Abstract of Israel* 53, Jerusalem: Central Bureau of Statistics, 2002.
- Donner, Allan, K. Stephen Brown, and Penny Brasher, "A Methodological Review of Non-Therapeutic Intervention Trials Employing Cluster Randomization, 1979-1989," *International Journal of Epidemiology* 19, 795-800.
- Dee, Thomas S., and Brian Jacob, "Do High School Exit Exams Influence Educational Attainment or Labor Market Performance," NBER Working Paper 12199, April 2006.
- 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, "Inference with Differences-in-Differences and Other Panel Data," *The Review of Economics and Statistics* 89 (May 2007), 221-233.
- Duckworth, Angela Lee and Martin P. Seligman, "Self-Discipline Gives Girls the Edge: Gender in Self-Discipline, Grades, and Achievement Test Scores." *Journal of Educational Psychology* 98 (2006), 198-208.
- Dynarski, Mark, and Philip Gleason, *How Can We Help? What Have We Learned from Evaluations of Federal Dropout-Prevention Program*, Princeton, NJ: MPR Report 8014-140, June, 1998.
- Dynarski, Susan, "Finishing College: The Role of State Policy in Degree Attainment," Harvard University Kennedy School of Government, mimeo, April 2005.
- 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.
- Frisch, R. and J. Moalem, "The Rise in the Return to Schooling in Israel in 1976-1997," Research Department, Bank of Israel, Working Paper No.99.06, 1999 (in Hebrew).

- Frisch, R., "The Return to Schooling - the Causal Link Between Schooling and Earnings," Research Department, Bank of Israel, Working Paper No.2007.03, 2007 (in Hebrew).
- Fuller, W.C., C.F. Manski, and D.A. Wise, "New Evidence on the Economic Determinants of Post-secondary 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.
- Garibaldi, Pietro, Francesco Giavazzi, Andrea Ichino, and Enrico Rettore, "College Cost and Time to Obtain a Degree: Evidence from Tuition Discontinuities," University of Turin, mimeo, July 2006.
- Gruber, J., "Risky Behavior Among Youths: An Economic Analysis," NBER Working Paper 7781 (July 2000).
- Hotz, V. Joseph, Lixin Colin Xu, Marta Tienda, and Avner Ahituv, "Are There Returns to the Wages of Young Men From Working While in School?," *The Review of Economics and Statistics* 84 (May 2002), 221-236.
- 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.
- Jackson, C. Kirabo, "A Little Now for a Lot Later: A Look at a Texas Advanced Placement Incentive Program," Harvard Department of Economics, mimeo, September 2007.
- 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.
- Kling, Jeffrey R., Jeffrey B. Leibman, and Lawrence F. Katz, "Experimental Analysis of Neighborhood Effects," NBER Working Paper 11577, August 2005.
- Kremer, Michael, Edward Miguel, and Rebecca Thornton, "Incentives to Learn," Harvard Department of Economics, mimeo, October 2003.
- Lalonde, Robert J., "The Promise of Public-Sector Training Programs," *Journal of Economic Perspectives* 9 (Spring 1995), 149-168.
- Leuven, E., H. Oosterbeek and B. van der Klaauw,, "The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment," CEPR Discussion Paper no. 3921, June 2003.
- Liang, Kung-ye, 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.
- Martorell, Francisco, "Do High School Graduation Exams Matter? Evaluating the Effects of Exit Exam Performance on Student Outcomes," mimeo, Berkeley Department of Economics, July 2005.
- 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.
- Medina, Jennifer, "Schools Plan to Pay Cash for Marks," *The New York Times* (June 19, 2007).
- 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).

- Schultz, T. Paul, "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program," *Journal of Development Economics* 74 (June), 199-250.
- Silverman, Irwin W., "Gender Differences in the Delay of Gratification: A Meta-Analysis," *Sex Roles* 49 (November 2003), 451-463.
- Stinebrickner, Todd R. And R. Stinebrickner, "Working During School and Academic Performance," *Journal of Labor Economics* 21 (April 2003), 473-491.
- Stinebrickner, Todd R. And R. Stinebrickner, "Time-use and College Outcomes," *Journal of Econometrics* 121 (2004), 243-269.
- Stinebrickner, Todd R. And R. Stinebrickner, "The Causal Effect of Studying on Academic Performance," NBER Working Paper 13341, August 2007.
- 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.
- Tyler, John H., "Using State Child Labor Laws to Identify the Effect of School-Year Work on High School Achievement," *Journal of Labor Economics* 21 (2003), 381-408.
- Warner, John T. and Saul Pleeter, "The Personal Discount Rate: Evidence from Military Downsizing Programs," *The American Economic Review* 91 (March 2001), 33-53.
- Wooldridge, Jeffery M., "Cluster Sample Methods in Applied Econometrics," *The American Economic Review* 93 (May 2003), 133-138.