Appendix Material for Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs

Jonathan M.V. Davis, University of Chicago
Sara B. Heller, University of Pennsylvania and NBER

May 17, 2017

A Youth Employment Program Literature

The research on active labor market programs has been reviewed in a number of other places (Stanley et al., 1998; Heckman et al., 1999; LaLonde, 2003; Card et al., 2015; Crépon and van den Berg, 2016), so we focus on only a few key lessons about youth programs here. Different authors draw different conclusions about whether employment programs for youth in particular are worth additional investment (Heckman and Krueger, 2004; Heinrich and Holzer, 2011; LaLonde, 2003; Raphael, 2012), but most point out that few well-evaluated employment interventions improve youths' labor market outcomes.

It is not the case that "nothing works." But the programs that do improve youths' labor market outcomes – programs like Job Corps, the National Guard ChalleNGe, and Year Up – involve lengthy and intensive interventions of a year or more, often with a residential component (Millenky et al., 2011; Roder and Elliott, 2011; Schochet et al., 2008). Benefit-cost analyses conducted close to program completion, which by necessity extrapolate short-term gains into the future, do sometimes suggest that the programs have positive returns (Mcconnell and Glazerman, 2001; Perez-Arce et al., 2012). But longer-term analyses that capture effect fade-out suggest that the returns to youth employment programs are, on average, more discouraging. Job Corps, which usually involves an average residential stay of about 8 months, has costs higher than the benefits it generates (Schochet et al., 2008). The National Guard ChalleNGe, a pseudo-military education program with residential and non-residential components lasting more than a year, may generate benefits in excess of costs, though it depends how educational benefits play out over time (Perez-Arce et al., 2012). Year Up, which provides 6 months of classroom training followed by a 6-month internship, increases wages but not employment rates in the year after the program (Roder and Elliott, 2011). To our knowledge, there is not yet a cost-benefit analysis.

Much of the employment literature ignores effects on other outcomes like crime (Crépon and van den Berg, 2016). But among the large, well-evaluated youth employment programs that measure criminal behavior, effects are even more mixed than for employment. Only Job Corps and JobSTART reduce crime (Cave et al., 1993; Schochet et al., 2008) (whereas the National Guard ChalleNGe has no effect on crime, and the Job Training Partnership Act may actually increase

crime among male youth) (Bloom et al., 1997; Millenky et al., 2011). The Job Corps and Job-START crime reductions, however, fade out quickly after the end of the programs, raising the possibility that only intensive programs reduce crime because their effects are limited to incapacitation during the program itself. The National Supported Work Demonstration also appears to have reduced crime among older participants, but not among youth (Uggen, 2000).

There is also a small, recent experimental literature on summer jobs programs in particular.[1] Leos-Urbel (2014) and Schwartz et al. (2015) use the allocation of program slots by lottery in New York City's summer jobs program to estimate program effects on school outcomes. They find small positive effects on attendance and test-taking among youth who are enrolled in school, with bigger effects for those who participate multiple times. Gelber et al. (2016) find that NYC's program reduces mortality by almost 20 percent and reduces incarceration for offenses committed as an adult by about 10 percent. They find no improvement in future employment and a small (~$100 per year) decrease in future wages. Using a large number of one-way interaction effects on demographic characteristics and employment history, they identify some treatment heterogeneity consistent with what we find here: better employment effects for younger, Hispanic youth. With their one-way interaction effects, however, they only find youth who do not have negative employment effects; no one in their subgroup tests seems to benefit on employment outcomes. They also find larger incarceration effects for the part of their sample that most resembles our study population (disadvantaged youth), and those at the highest risk of incarceration (minorities and males).

We know of several unpublished works-in-progress on summer jobs programs as well: Gelber et al. are analyzing additional crime data, and Alicia Modestino is studying Boston's summer jobs program. One of us (Heller) is also working on other studies of Chicago summer programs and a study of Philadelphia's WorkReady program.

B Data

This section provides additional details about our data sources and variable definitions.

B.1 Crime Data

Our crime data are drawn from administrative arrest records provided by the Illinois State Police (ISP). The data capture arrests in the state of Illinois since 1990, before any study youth were born. Initial coverage was somewhat incomplete, but coverage improved considerably over the 1990s. Individuals are identified in the ISP system using fingerprint cards which must be submitted for all felonies and class A and B misdemeanors and may be submitted for Class C misdemeanors.

We determine the crime associated with every charge by parsing the description of the crime. When an arrest is associated with several charges, we restrict attention to only the most serious charge. We then aggregate the crime categories so that each arrest is classified as either a violent, property, drug, or other crime. The specific subcategories for each crime are:

- Violent Crime: aggravated arson (when a person was known to be home), aggravated assault,

[1] The appendix to Heller (2014) summarizes the less recent summer employment literature. In addition to some non-experimental evidence, this includes a study of STEP that used random assignment but provided training and jobs to both treatment and control groups, estimating only the added value of a life skills and sex education curriculum (Grossman et al., 1987; Grossman and Sipe, 1992; Walker and Vilella-Velez, 1992), and a study in Philadelphia that used random assignment but analyzed the data using non-experimental methods (McClanahan et al., 2004).

aggravated hijackings, armed robbery, assault, home invasion resulting in injury, homicide, intimidation, kidnapping, making a threat, robbery, sex offenses, and violating an order of protection;

- Property Crime: arson, burglary, counterfeiting, deceptive practices, home invasion without injury, identity theft, larceny, and motor vehicle theft;
- Drug Crime: drug dealing or possession;
- Other Crime: bribery, child endangerment, contributing to delinquency of a minor, dog fighting, driving violations (including DUIs), disorderly conduct, dumping, fleeing police, gambling, indecent exposure, obstruction of justice, ordinance violations, parole violations, prostitution, reckless conduct, resisting a police officer, trespassing, underage alcohol consumption, warrant for arrest, weapons charges, and vandalism.

ISP creates the arrest extract by matching study youth to arrest records using the following fuzzy matching procedure: First, youth are deterministically matched to arrest records based on their first and last name and date of birth. Exact matches are added to the extract. To account for typographical errors, youth are then probabilistically matched to arrest records using the MergeToolBox (MTB) software package. For feasibility, the set of potential matches is limited to observations where either the first name, last name, or date of birth match exactly. Potential matches are evaluated using MTB's scoring procedure which takes the average string similarity between first and last names as measured by the Jaro-Winkler algorithm and dates of birth as measured by the Damerau-Levenshtein distance as inputs. We did not share youths' treatment statuses with ISP. The baseline balance tests demonstrate that, as expected, there is not a statistically significant difference in the probability that treatment and control group youth were matched to a pre-randomization arrest record. Since we observe the date of arrest, we define pre-randomization arrests based on the date of the lottery used to select a youth for the treatment or control group. An arrest is considered pre-randomization if it occurred before the date of the lottery.

B.2 School Data

Our school data are drawn from administrative records provided by Chicago Public Schools (CPS). The data include details on each student's demographics, enrollment, attendance, and grades from the 2011-12 school year through the end of the 2015-16 school year.

The research team matched study youth to CPS's master enrollment file, which includes all youth who have enrolled in a CPS school since 1988. Because the 2012 cohort was recruited through schools, they all appear in the school records. In the 2013 cohort, 91.66% of youth were successfully linked to a CPS record, with no difference in the matching rate between treatment and control groups (p=0.672). We matched youth using historical CPS records beginning in 1988, before the oldest youth in the study was born, so that any youth who had ever enrolled in CPS would be matched to their schooling records. There are two reasons that study youth might not be found in the CPS data. First, they may not have ever attended school in Chicago. About 1% of the sample lived outside the school district limits (but still in Cook County) at the time of their application. These youth would not be in the CPS data unless they had previously lived in the city and enrolled in CPS during prior years. Others may have attended only private or parochial schools. Since youth are legally allowed to drop out of school at age 17, some youth may also

have been above the age of legally-required school enrollment when they moved into the district. Second, data errors may have generated false negatives in the matching process.

Sample observations were matched to administrative CPS records using MTB based on their first and last names and dates of birth. Treatment status was never considered in the matching procedure. As expected, there is not a significant difference in the probability that treatment and control group youth are matched to a CPS record. In total, 435 youth are not matched to a CPS record either because they never attended a CPS school or because data errors prevented matching. All schooling variables are necessarily missing for these observations.

The master enrollment file includes details on youth's enrollment status, exit reason (i.e. graduation, transfer, etc.) and date, grade level, free or reduced price lunch status, disability status, and demographics. Students are then matched to separate attendance and grade files using their CPS identifier.

We define several baseline covariates using youth's schooling information from the school year prior to randomization. These baseline characteristics are missing for the 29% of youth matched to a CPS record who did not attend a CPS school in the pre-program school year (i.e., had already graduated, transferred, or dropped out).

The attendance records include counts of days present, absent, and enrolled at every school a student attended each year. We aggregate these counts to the student level by adding across schools. Our baseline attendance measures are missing for 5 students who enrolled in school in the pre-program school year.

We also use the attendance records to generate our indicators for the type of each student's main school of attendance. CPS classifies its high schools by type and enrollment style. We identify a student's school classification as the the classification of school at which the student attended the most days in the pre-program school year. In post-program school years, we define enrollment based on whether a student attended at least one day of school. When measuring attendance as an outcome, we restrict attention to attendance at non-detention center schools (when youth are incarcerated in detention or prison, they must attend the schools in those facilities).

The grade file includes students' grades for each term. We use course grades from the pre-program fall semester (call this $t-1$) to generate student i's baseline GPA according to the following formula:

$$GPA_{i,t-1} = \frac{4*\#A_{i,t-1}+3*\#B_{i,t-1}+2*\#C_{i,t-1}+\#D_{i,t-1}}{\#A_{i,t-1}+\#B_{i,t-1}+\#C_{i,t-1}+\#D_{i,t-1}+\#F_{i,t-1}}.$$

We exclude spring semester grades from our baseline GPA measure because some spring grades were finalized post-randomization. Baseline GPA is missing for 1,206 youth who attended a CPS school in the pre-program school year. GPAs are generally missing because the youth attended charter or other non-traditional schools that did not report grades using CPS's main system, or because the youth attended very few days of school. GPA is missing for 790 (25%) youth who attended at least 1 day of school in the first post-program school year.

We define our GPA outcomes analogously to our baseline measure only using grades in both the fall and spring semesters of the relevant post-randomization school year at non-detention center schools. There are 138 youth who only received grades at a detention center school.

B.3 Employment Data

Our employment data come from administrative records provided by the Illinois Department of Employment Security (IDES). The data provide a quarterly record of all UI-covered jobs held by study youth between 2005Q1 and 2015Q1. For every job-quarter, the data report the employer name, industry, and total wages.

We define a youth as having any formal employment if they appear in the employment data in a particular quarter. We do not require youth to have any earnings to count them as employed. This is because some of the One Summer Plus employment providers reported youth as working but with no earnings (treating program wages as stipends). We use employer names to identify whether youth were working for One Summer Plus employment providers or other employers.

We obtained youths' employment records through a joint data sharing agreement between CPS and IDES. CPS no longer asks youth for their social security number when they enroll. However, most youth in the study enrolled in CPS under the old policy which often collected an SSN at registration. Therefore, CPS's internal records (not shared with the research team) include most study youths' SSNs. CPS sent IDES a list of youth's social security numbers, which IDES used to create the extracts.

In total, we have employment data on 5,076 youth (3,426 of whom actually have an employment record; 1,650 were never employed). Of the 1,774 youth for whom CPS did not have a SSN available for matching, 435 are missing because they were not matched to a CPS record and 1,339 did not have a valid SSN in their CPS record.

B.4 Subsample Baseline Balance

The main text shows descriptive statistics and baseline balance for each cohort (2012 and 2013) separately. Table A1 shows the same descriptive statistics for the pooled sample, while Tables A2 and A3 describe the samples with non-missing school and employment data respectively. Table A4 describes the sample used for our main schooling analysis, which includes observations with a CPS record who had not graduated prior to the program. The subsamples are generally fairly similar to the sample as a whole, and all pairwise and joint tests suggest that the treatment and control groups are balanced in each subsample.

C Additional 2013 Randomization Details

Details on the 2012 experimental design are in Heller (2014). In the 2013 study, youth were recruited from two applicant pools. The first group was recruited directly from the justice system. The second group was applicants to Chicago's broader One Summer Chicago (OSC) program living in neighborhoods with the highest violent crime rates.

The web application for OSC through which youth in the second applicant pool applied was not designed solely for OSC+, and it did not ask applicants to report their gender. Since the 2013 program was for boys only, we randomly assigned all applicants in this second pool (n = 7,588) and relied on program providers to discern gender (those assigned to treatment were contacted by the service providers and not offered the program if female, with the exception of a very small number of transgender youth who were female but identified as male and so were allowed to participate). After the lottery, we matched youth to other administrative data sources (school and arrest records) that include gender. The analyses reported here drop female applicants. This does not undermine the integrity of random assignment, since gender is a baseline characteristic. We do not observe

gender for the observations that were not located in schooling records and had no pre-program arrests (n=351). We include these observations in the analysis, with an indicator variable for the fact that they are missing gender included as a covariate.

Youth were blocked on applicant pool and age (under or over 18), largely because the city had a legal obligation to keep probationers who were under age 18 separate from those who were over 18. In order to ensure that each of the service providers were assigned the number of youth they were able to serve, and to minimize the distance youth had to travel to the provider offices, we also blocked on the geographic location of applicants' home address.

Both because of the abbreviated recruitment time frame and the missing gender issue, we provided the program providers with a much longer list of treatment group youth than could actually be served. While names were randomly sorted on the lists, program providers could have potentially used discretion when determining the order in which they contacted treatment group youth about the program. As would be expected if program providers just worked from the top of the lists, youth's rank on the treatment list is negatively associated with participating in the program. If we regress an indicator for participation on treatment status, youth's rank on a providers' sheet, our baseline covariates, and block fixed effects, the coefficient on the youth's rank is -0.00014 (SE=0.000045, $p<0.01$), though adding list rank as an instrument does not appreciably improve the first stage. Control group names were never shared with the providers.

Despite a de-duplication process at the time of application, 52 youth submitted duplicate applications that were not identified until after random assignment. We consider a youth to be in the treatment group if any of his applications were assigned to it, effectively using the maximum of a youth's random assignment indicators as their final assignment. This ensures that treatment assignment is still random, since the maximum of any series of random variables is also random. For the analysis, we only retain one observation per youth. Because those who entered the lottery more than once had a higher probability of being selected, treatment assignment is only random conditional on the number of applications submitted. To account for this, we include indicator variables for submitting one or two duplicate applications in all of our analyses.

D Multiple Hypothesis Testing Adjustments

D.1 Permutation Tests and the Familywise Error Rate (FWER)

Following Westfall and Young (1993) and Anderson (2008), we adjust our inference in several ways to relax assumptions and account for the number of hypothesis tests we are conducting. First, we calculate a "randomization" or "permutation" p-value (Lehman and Romano, 2005). This avoids the need to rely on asymptotics or distributional assumptions, and instead calculates p-values based on the empirical distribution of test statistics under the null of no treatment effects. The steps are as follows:

1. Enforce the null of no treatment effect by re-randomizing treatment assignment in the data using the same sampling frame as in the initial random assignment.

2. Using the new pseudo-treatment assignment, calculate and save the t-statistic for each of the hypothesis tests.

3. Repeat this process 100,000 times.

4. For each hypothesis test, compare the actual t-statistic in the sample with the empirical distribution of t-statistics generated under the null of no treatment effect. (For two-sided tests, use the absolute value of all t-statistics). Let the permutation p-value equal the probability that a simultated t-statistic is greater than the obtained t-statistic.

5. Reject the null if this probability is less than some critical value α.

The Familywise Error Rate (FWER) adjustment uses a similar re-sampling framework, but adjusts for the number of tests conducted within a family of hypothesis tests, $\mathscr{F}$. A test controls the FWER at a level α if the probability of falsely rejecting at least one hypothesis in the family $\mathscr{F}$ is no greater than α.

We calculate FWER adjusted p-values using a variant of the step-down procedure described by Anderson (2008). We start with the set of t-statistics generated under the null of no treatment effect across 100,000 permutations from steps 1-3 above. We define a set of related outcomes as a family $\mathscr{F}$, where each hypothesis test within the family is associated with a test statistic t. We use the absolute value of the t-statistic on the treatment effect from our standard regressions as our test statistic.

The step-down procedure for adjusting p-values to control the FWER drops a hypothesis from the family once it has been adjusted to increase power on the remaining tests. The steps are as follows:

1. In the original sample, consider the set of f hypothesis tests that are part of family $\mathscr{F}$. Sort the test statistics in that family from greatest to least (initially, $t_1 > t_2 > \cdots > t_f$). Let r index the different outcomes within the family by the rank of the associated test statistic (so the largest t statistic in family $\mathscr{F}$ has r=1, the next largest has r=2, and so on).

2. Within each permutation, choose the largest t in family $\mathscr{F}$ (even if it is associated with a different hypothesis test than t_1). Construct a distribution of these maximum test statistics in the family.

3. Calculate the proportion of this distribution that is at least as large as t_1. This proportion is the initial FWER-adjusted p-value for outcome 1's hypothesis test. Call this p_1^* - the probability of falsely rejecting that hypothesis given all the tests in the family.

4. In both the original sample and the permutations, drop the test statistic associated with outcome 1, so $\mathscr{F}$ now contains outcomes 2 through f. Repeat steps 1-4 with the remaining tests until there are no tests left in the family.

5. Ensure the FWER adjusted p-values have the same rank ordering as the original test statistics (i.e., ensure $p_1^{FWER} < p_2^{FWER}$ so that the larger unadjusted p-value is always associated with the larger adjusted p-value). Note in the initial ordering in step 1, $p_1 < p_2 < ... < p_f$. To enforce this same ranking within the adjusted p-values, replace each adjusted p-value with the maximum of the adjusted p-values of all the tests with smaller ranks, inclusive: $p_r^{FWER} = max\{p_1^*, p_2^*, \ldots, p_r^*\}$.

D.2 False Discovery Rate (FDR)

The False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) is the expected proportion of rejected hypotheses in a family of tests which are false rejections. Using Anderson's (2008) implementation of Benjamini and Hochberg's one step procedure, we calculate q-values, which are the FDR's version of a p-value (the smallest proportion of false rejections we could accept in family $\mathscr{F}$ and still reject the hypothesis).

To decide whether to reject the null of no treatment effect while controlling the FDR at level q, order the p-values associated with the f hypothesis tests in family $\mathscr{F}$ from smallest to largest, such that $p_1 < \ldots < p_f$. Let f equal the number of hypotheses in the family, and let r denote the rank of a particular test in this ordering (so as above, the smallest p-value in family $\mathscr{F}$ has r=1, the next smallest has r=2, and so on). Find the largest rank $r*$ for which $p_r \leq {}^{rq}/_{f}$. Reject the null for all hypotheses with rank from 1 to $r*$.

The q-values reported in Table A5 are the smallest values of q for which each hypothesis would be rejected according to this procedure.

D.3 Adjusted Results

Table A5 shows the main results with the previously described adjustments for multiple hypothesis testing. Each panel of the table shows the results for a different family of outcomes. Panels A and B show the adjustments for arrests in year one and two, respectively; panels C and D show the adjustments on employment outcomes during the program and in the post-program quarters, respectively; and panel E shows the adjustments for our main schooling outcomes. The first two columns of the table show the control complier mean (CCM) and LATE for each outcome. The remaining columns show four different versions of p-values. First, we show the standard p-value for a single two-sided test. Next, we show the "permutation" or "randomization" p-value, which is the probability of observing a t-statistic (in absolute value) at least as extreme as the one in our data across 100,000 permutations of treatment assignment. Third, we show the q-value from Benjamini and Hochberg's (1995) procedure to control the False Discovery Rate (FDR), using the p-values in column 3 as inputs into the procedure. This reports the smallest level of q at which the null hypothesis would be rejected (where q is the expected proportion of false rejections within the family, or the level at which the false discovery rate is controlled). Finally, we show p-values which control the Familywise Error Rate. Across the adjustments, the year-one reduction in violent-crime arrests, the increase in employment during the program quarters, and the increase in employment at program providers after the program remain statistically significant.

E Causal Forest Details

In a world of heterogeneous treatment effects, policymakers must take care when scaling up "evidence-based" programs. If not all youth benefit in the same way, expanding the program to different populations might not replicate the documented program effect. We varied OSC+'s recruitment strategies across two cohorts in order to explore heterogeneous treatment effects across different types of youth, but we want to guard our analysis against the false positives that can result from testing program effects across many subgroups, as well as gain a more nuanced understanding of who benefits than is possible with researcher-specified one- or two-way interaction effects.

To do so, we use the causal forest approach of Wager and Athey (2015). The causal forest adapts regression tree algorithms (Morgan and Sonquist, 1963; Morgan, 2005; Loh, 2014) from the

supervised machine learning literature to the problem of causal inference. The following section provides some background on regression trees and random forests, as well as how Wager and Athey (2015) adapt these methods to predict causal effects, in order to provide some intuition about the approach. It then explains the details of how we implement the procedure. Some of the material below builds on the shorter explanation in Davis and Heller (2017).

E.1 Regression Trees and Random Forests

This section provides a very brief primer on how regression trees and random forests work when the aim is to predict an outcome variable, Y. There are many more complete reviews of this part of the machine learning literature (Breiman et al., 1984; Hastie et al., 2009; James et al., 2013), but our hope is to provide just enough intuition to allow a reader unfamiliar with these methods to understand the basic underpinnings of the causal forest algorithm we use in the main text.

A regression tree builds a potentially complex non-linear model of Y as a function of a large set of covariates. To do so, it recursively partitions the data a single covariate at a time (e.g., start by splitting into male and female, then split each gender into income groups, and so on). The result is a series of splits in the data, defined by values of covariates, where each group formed by a split is a "leaf" of the tree, and successive splits grow out of the prior leaf.

To construct, or "grow," the tree, an algorithm starts with the unpartioned data (all the observations, usually of a subset of the data called the training data set). It then searches over all the possible splits of the data based on a single covariate. Each potential split forms two leaves branching off the parent node.[2] Potential splits are generally evaluated by their predictive power, i.e., by minimizing an in-sample goodness-of-fit criterion like Mean Squared Error = $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$, where $\hat{y}_i$ is the mean of y in i's leaf. The split that generates the best fit (e.g., smallest mean squared error) is kept, generating 2 child leaves branching off the parent node. Then the algorithm searches over the possible splits in each of the two new leaves, and so on. Trees are often built with "greedy" algorithms, which start at the top (no splits) and consider the next split based only on what maximizes the fit at that step, ignoring any implications for splits that will eventually follow below.[3] The algorithm stops splitting when a stopping rule is reached (there are fewer than a specified number of observations left in a leaf after a potential split, or additional splits do not improve the goodness-of-fit).

The subgroups formed by the last set of splits, where the tree ends, are called terminal nodes or terminal leaves. These terminal leaves divide the covariate space into non-overlapping regions. Formally, let $l(x;\Pi)$ be the terminal leaf of a tree Π that contains observations with a vector of covariates $X_i = x$. A tree is a collection of these terminal leaves: $\Pi = \{l_1, \ldots, l_{\#(\Pi)}\}$ where the collection of leaves covers the full covariate space. The tree can then provide out-of-sample predictions of Y based on Xs by figuring out to which terminal node an observation belongs based on its covariates and assigning the mean Y within that leaf to the observation. This prediction strategy is a version of nearest-neighbor matching, where the neighborhood is the leaf, which is determined by the tree-growing process rather than a pre-defined distance measure.

Using those predictions directly, however, would suffer from over-fitting and likely perform

[2]To cement ideas, if the covariate X_1 is a dummy variable, the algorithm considers splitting the data into observations where X_1= 0 and those with X_1=1. If it is continuous, it considers a split at all possible values of X_1.

[3]This generally means that a single tree is not necessarily "optimal," insofar as the algorithm will always make the split that minimizes the mean squared error at a particular step, even if a different split would result in better predictive accuracy farther down the tree.

badly on out-of-sample data. One potentially better approach is to "prune" the tree by removing splits to make a smaller tree. In order to decide which splits to prune, one can adjust the goodness-of-fit criteria to include a penalty for complexity (i.e., max $[-\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 - \alpha|T|]$, where $|T|$ is the number of nodes and α is a tuning parameter chosen using cross-validation to maximize the out-of-sample predictive accuracy of the model). This approach will remove the splits where the benefit of the split in terms of mean-squared error reduction does not outweigh the cost of having a more complex model.

However, using a single tree for prediction in this way may not always be desirable; it is a high variance approach with no guarantee that a given tree is optimal. A different approach is to grow many trees and average across them. In a process called bootstrap aggregating, or "bagging," hundreds or thousands of random subsamples of the data are selected, and a tree is grown on each subsample (called the "estimation sample"). The subsamples should be small relative to the data in order to de-correlate the estimates across trees.[4] Predictions for any given individual are assigned as the average of the predictions across all the trees for that individual. Bagging helps to reduce bias by allowing deep (unpruned) trees that narrow the neighborhood represented by each leaf, and to reduce variance - without requiring any cross-validation or leaf penalty - by averaging across many predictions (Breiman, 1996; James et al., 2013). The variance of the predictions can be calculated from the "out-of-bag" observations, or the variation in the predictions across trees when an observation was not part of the estimation sample.

A "random forest" is very close to bagging, with one small adjustment. At each split, only a random subset of covariates is considered when selecting the best split. This prevents a few strong predictors from being used over and over again across trees, which helps to decorrelate the trees and reduce prediction error.

E.2 Causal Forest

Wager and Athey's (2015) causal forest method relies on many of the same principles as regression trees and random forests, but aims to predict treatment effects rather than outcomes. Predicting treatment effects is more difficult than predicting outcomes because treatment effects are fundamentally unobservable at the individual level (making it impossible to calculate a mean squared error directly). Instead, the researcher observes:

$$Y_i^{obs} = Z_i Y_i(1) + (1 - Z_i) Y_i(0),$$

where $(Y_i(1), Y_i(0))$ are potential outcomes for individual i. Specifically, $Y_i(1)$ and $Y_i(0)$ are individual i's outcomes when treated or not treated, respectively. Z_i is an indicator for treatment status which, in this section, we assume is assigned completely at random.

A Conditional Average Treatment Effect or CATE is the causal effect of a treatment for a subgroup defined by covariates:

$$\tau(x) = E[Y_i(1) - Y_i(0) | X_i = x] \tag{1}$$

where Y is the outcome of interest, X is a vector of observable baseline covariates, and x is a particular realization of the baseline covariates. If $\tau(x)$ is not constant for all Xs, policymakers

[4] Formally, the ratio of the subsample size relative to the sample size should converge to 0 as the sample size goes to infinity (Romano and Shaikh, 2008).

would likely want to target the intervention to those subsets of the population with the largest (or smallest) CATEs.[5]

With a large enough dataset, provided that treatment assignment is orthogonal to potential outcomes conditional on X (the "unconfoundedness" assumption), we could estimate conditional average treatment effects (CATEs) with the difference of treatment and control means for every unique subgroup in the data, then investigate treatment heterogeneity across these estimates. In practice, this is not feasible since the number of unique subgroups grows large very quickly with the number of discrete covariates or with even a single continuous covariate. For example, with 10 binary covariates, there are 1024 (2^{10}) unique subgroups defined by different covariate realizations. Not only would we need an infeasibly large sample to estimate these effects, but the number of hypothesis tests required to test differences across these subgroups would also generate too many false positives to be useful.

Causal forests address these problems by adapting the classification and regression tree (CART) methodology to the problem of estimating $\tau(x)$. They provide a feasible approximation to the above procedure by attempting to include only "important" covariates in the conditioning set, where importance is determined by how much a goodness of fit measure improves with a covariate's inclusion. Traditional mean squared error measures of goodness of fit cannot be used directly, since treatment effects are not observed for any individual (i.e., we do not observe the "ground truth" for an individual, or $\tau_i(x)$, by which to measure how well a prediction does in an individual case). Athey and Imbens' (2016) causal tree algorithm adapts the CART approach to the estimation of conditional average treatment effects using several novel in- and out-of-sample goodness-of-fit measures. In particular, they show that selecting splits in order to maximize heterogeneity in treatment effects across leaves, less a penalty for the variance of treatment and control outcomes in each leaf, is equivalent to using expected mean squared error.[6]

Specifically, they propose choosing splits at each leaf using a greedy algorithm that maximizes the following objective function:

$$\frac{1}{N}\sum[n_l\hat{\tau}_l^2 - 2(\frac{\hat{Var}(Y_{l,treat})}{p_l} + \frac{\hat{Var}(Y_{l,control})}{1-p_l})], \tag{2}$$

where N is the number of observations used to estimate the tree Π; n_l is the number of observations assigned to leaf l, p_l is the proportion of treatment observations in leaf l; $\hat{Var}(Y_{l,treat})$ and $\hat{Var}(Y_{l,control})$ are the variances of treatment and control outcomes among observations in leaf l, respectively; and $\hat{\tau}_l$ is the estimated conditional average treatment effect in leaf l. Provided the unconfoundedness assumption holds, this conditional average treatment effect can be estimated as the difference between treatment and control mean outcomes in each terminal leaf. That is, within

[5]There are of course reasons this might not be true. If the cost of serving different types of people varies, one would also want to consider the costs, not just the benefits, of serving a particular subgroup. If the stable unit treatment variance assumption does not hold (e.g., a group's CATE depends on who else they interact with in the program), targeting only the groups with the highest CATEs might have other general equilibrium effects that could make such targeting suboptimal. And if policymakers value equity or prefer particular distributional consequences, targeting those with the largest gains may not achieve their goals. Nonetheless, learning who benefits the most from one particular implementation of an intervention is still a useful way to form hypotheses about the consequences of scale-up or targeting changes, which could then be tested in practice and weighed against policymakers' preferences.

[6]Athey and Imbens (2016) show that maximizing the variance of treatment effects is equivalent to maximizing $-\sum_{i=1}^{n}(Y_i(1) - Y_i(0) - \hat{\tau}_i)^2$.

each leaf l:

$$\hat{\tau}_l = \bar{y}_{l,treat} - \bar{y}_{l,control}.$$

As is standard in bagging over regression trees (James et al., 2013), causal trees are estimated with no pruning.

Causal forests avoid relying on any single tree by assigning individual observations the average of their predicted effects across a large number of trees estimated on random subsamples of the data drawn without replacement. As with random forests, we further "de-correlate" trees by considering only a random subsample of covariates when determining each split.

Wager and Athey (2015) recommend determining the structure of the trees (defining the l's) and estimating the treatment effects within leaves (the $\hat{\tau}_l$'s) using separate subsamples. Trees which are built and estimated with independent samples are called "honest." In practice, this means splitting each random subsample into two, using half the observations to grow the tree and the other half to estimate the treatment effects in each leaf. In our companion paper, we show that "honest" causal forests are prone to overfitting if we assign $\hat{\tau}_l$ to all observations in the sample including those used to grow the tree and estimate the $\hat{\tau}_l$'s (Davis and Heller, 2017). Instead, we average predicted treatment effects only across the predictions made when an observation is neither part of the tree-growing subsample nor part of the subsample used to calculate $\hat{\tau}_l$.

With conditional average treatment effects based on each individual's covariates in hand, researchers can examine who is expected to respond most to the treatment based on their observable characteristics. This approach to estimating treatment heterogeneity allows covariates to matter in a much more flexible way than would be possible using a small subset of pre-specified interaction effects, and it avoids over-fitting and spurious effects by using our adjusted "honest" approach. Another benefit is that it predicts an entire distribution of conditional average treatment effects, rather than just a handful of subgroup effects.

To make the process concrete, we provide step by step instructions for how we implement the causal forest in our companion paper (Davis and Heller, 2017).[7] As discussed in that paper, the researcher is responsible for choosing three parameters: (1) the number of trees in the forest; (2) the minimum number of treatment and control observations allowed in a leaf; and (3) the subsample size. Increasing the number of trees in the forest reduces the Monte Carlo error due to randomly selecting subsets of the data when estimating the causal forest. Therefore, the researcher should select the number of trees to be as large as is possible given computational constraints. In our experimentation with these parameters, we found that our estimates were more stable using 100,000 than 25,000 trees. The choice of the minimum number of treatment and control observations required in each leaf is less straightforward because there is a bias-variance tradeoff. Bigger minimum leaf sizes reduce variance but increase bias. There is similarly a tradeoff when selecting the subsample size. Bigger subsamples allow for bigger and more precisely estimated trees, but increase the correlation of the trees' estimates across subsamples.

In our setting, we found little difference between estimates from a forest estimated using 10 percent subsamples of the data and those of a forest estimated with 100% bootstrap samples drawn with replacement. The estimates reported in the main text use a 20 percent subsample, split evenly between tree-growing and estimation subsamples. We did, however, find that there is a trade-off in

[7] We are deeply indebted to Susan Athey for providing a beta version of the causal forest code, which we adapted for our analysis.

the choice of leaf size; bigger minimum leaf sizes improve the stability of the predictions across different samples but reduces the amount of heterogeneity the forest identifies. We require all leaves to have at least 10 treatment and 10 control observations.

E.3 Implementation Details in Our Context

We estimate causal forests using a large set of policy relevant covariates measured at or before random assignment. We avoid using variables that are only available for a subset of data (e.g., prior-year GPA is only available for students in the CPS data who were still in school prior to the program, and prior-year wages are only available for those with valid SSNs who worked). Instead, we define covariates that are available for everyone as follows:

- Demographic: Age in years and indicator variables for being male, Black, or Hispanic
- Neighborhood (from the ACS): Census tract unemployment rate, median income, proportion with at least a high school diploma, and proportion who rents their home;
- Crime: Number of pre-randomization arrests for violent crime, property crime, drug crime, and other crime;
- Education: Indicator variables for having graduated from CPS prior to the program, being enrolled in CPS in the school year prior to the program, not being enrolled in the year prior to the program despite having a prior CPS record, and not being in the CPS data at all;
- Employment: Indicator variables for having worked in the year prior to the quarter of randomization, for having not worked in the year prior to the quarter of randomization despite having a valid SSN, and for not having a valid SSN.

We observe all of these covariates for everyone in the sample except for gender, which is missing for 351 observations. We impute these observations using block means.

In terms of outcome variables, we observe crime outcomes for everyone but have to deal with missing outcome data for other measures. For employment, we restrict the causal forest sample to the 5,076 observations with a social security number (i.e., not missing employment data). For school persistence, we restrict the sample to the 6,415 youth who were matched to a CPS record.[8] We run the entire causal forest procedure using just observations with non-missing data for these outcomes.

An important assumption for the causal forest to produce consistent estimates of treatment effects is that within each leaf, treatment assignment is orthogonal to potential outcomes conditional on X. For this to be true in our case, we must condition on randomization block, since treatment probabilities varied across blocks. We adjust for differences in treatment probabilities using inverse probability weights (Athey and Imbens, 2016; Rosenbaum and Rubin, 1983; DiNardo et al., 1996). Specifically, we weight each observation by $w_i = \left(\frac{Z_i}{p_{b(i)}} + \frac{1-Z_i}{1-p_{b(i)}}\right)$ where Z_i is an indicator for being randomly assigned to the treatment group and $p_{b(i)}$ is the probability of being treated in observation i's block b, $p_{b(i)} = E(Z_i|B = b(i))$. These weights are used throughout the causal forest procedure. For example, the predicted treatment effect within a leaf is given by:

[8]The results are similar if we exclude pre-program graduates when we estimate the causal forest.

$$\hat{\tau}_l(x;\Pi) = \frac{\sum_{i\in l} Z_i w_i Y_i}{\sum_{i\in l} Z_i w_i} - \frac{\sum_{i\in l}(1-Z_i) w_i Y_i}{\sum_{i\in l}(1-Z_i) w_i}$$

The variance of outcomes by treatment status within a leaf, used as part of the algorithm's objective function above, is similarly weighted. This transformation effectively controls for differences in treatment probabilities across blocks by re-weighting the number of treatment and control observations to equal the total size of the block. Consider a hypothetical block with 25 treatment observations and 75 control observations. In this block, $p_b = 0.25$. Each treatment observation receives a weight of 4 and each control observation gets a weight of 4/3. With these weights, there are effectively 100 treatment (4*25) and 100 control (4/3*75) observations. This adjustment eliminates any differences in treatment probabilities across blocks while keeping the relative size of blocks the same.

E.4 Density of Predictions

Figures A1, A2, and A3 show the densities of the causal forest's predicted impacts on any post-program formal employment, violent crime arrests in the 2 or 3 years after the program (depending on the program year), and persistence in school through the third post-program school year. The distributions are fairly symmetric and bell-shaped, which contributed to our decision to use a high-quartile comparison to test heterogeneity in the main text. The bulk of the violent-crime predictions are below zero, consistent with the significant overall decline in that outcome. The average predicted intent-to-treat impacts on post-program formal employment, cumulative violent crime arrests, and persistence through the third post-program school year are .01, -1.83, and -0.01, respectively.

The coefficient of variation, which equals the standard deviation over the absolute value of the average, is a dimensionless measure of spread. The coefficients of variation for post-program formal employment, year one violent crime arrests, and persistence through the third post-program school year are 3.80, 1.30, and 1.22, respectively. In other words, the causal forest predicts substantially more variation in conditional average effects for post-program employment than for violent crime arrests or school persistence. Of course, the average impacts on employment and schooling are close to zero, which will increase the coefficient of variation. A more robust measure of dispersion is the Quartile Coefficient of Dispersion, which equals: $|\frac{Q3-Q1}{Q3+Q1}|$. We see a similar pattern using this measure: 2.70, 0.78, and 1.15.

F Additional Results and Robustness Checks

F.1 Participation

Table A6 shows the participation details referenced in Section 5 of the main text. Panels A and B summarize participation for the 2012 and 2013 summer programs, respectively. Panel C summarizes participation for the 2013 extension programming.

F.2 Main Results with No Controls

In the main text, all of our regression results control for a rich set of baseline controls (listed in the Analytical Methods section). Tables A7 through A9 show the main results controlling only for

block fixed effects and indicators for having one or two duplicate applications in the lottery, which are required for treatment to be (conditionally) random.

F.3 Intent-to-Treat and Alternative Functional Form

In the main text, we focus on the effects of participation for youth who choose to comply with random assignment. Tables A10 through A12 show the intent-to-treat estimates and control means for the main crime, schooling, and employment results.

Our main crime estimates treat the number of arrests as a continuous variable. In reality, however, these dependent variables are counts. Table A13 shows that using Poisson regression (with robust standard errors to allow for over-dispersion) does not change the substantive findings. The average marginal effects from the Poisson regressions are very similar to the ITT results reported in Table A10.[9]

Similarly, our employment analysis uses linear probability models for having any formal, provider, or non-provider employment. Table A14 shows that our substantive findings are unchanged if we estimate average marginal effects using a probit.

F.4 By Treatment Arm (2012 only)

In the 2012 study, youth were randomly assigned to two different treatment groups: one that worked 5 hours a day and one that worked 3 hours per day with the 2 other hours spent engaging in a social-emotional learning (SEL) curriculum. The curriculum was sometimes offered at the worksite, but sometimes required additional travel. Both groups had the adult job mentor and were assigned similar types of jobs.

Tables A15 through A17 show crime, school, and employment results separately by treatment arm for the 2012 cohort (the 2013 cohort did not have two treatment groups - everyone received the SEL curriculum). Although some program providers kept separate records on participation at a work site versus at the SEL training, these records are not universally reliable. As such, we focus on the ITT rather than trying to instrument for participation in each type of activity separately.[10]

In general, we are under-powered to cleanly distinguish differences between the two treatment arms. Both the jobs-only and the jobs + SEL groups show a substantively large decline in year 1 violent-crime arrests (with p-values of 0.128 and 0.047 respectively). The point estimate for the SEL group is bigger (3.5 versus 2.8 fewer violent-crime arrests per 100 youth), but the standard errors are too large to reject the null of no difference between groups. The direction of the program effects for non-violent crimes, however, does not universally favor the SEL group. The increase in property crime arrests after the first year seems concentrated among the SEL group, which has a large and significant increase in year 2 property crimes (4.6 per 100 youth) that is marginally significantly different from the jobs-only group (p-value for the test of no difference = 0.09), though the cumulative increases are statistically indistinguishable across treatment arms. The point estimates on "other" arrests also suggest an increase in these minor crimes during years 1 and 3 for the SEL group relative to the jobs-only group (both years' other-crime effects are significantly differ-

[9]The Poisson estimates are generally slightly less precise, because we use a parsimonious set of covariates (age, gender, and an indicator for any baseline arrests in addition to randomization block and duplicate application dummies) to ensure convergence. We also set Stata to cycle through different maximization algorithms for the same reason.

[10]In the data we have, compliance by treatment arm was fairly similar. Take-up of the job itself was almost identical: 72.3 percent among jobs-only youth and 73 percent among jobs + SEL youth. Only 6 percent of youth assigned to the jobs-only treatment arm attended any SEL sessions, whereas 63 percent of jobs + SEL youth attended.

ent across treatment groups, p= 0.08 in year 1 and 0.04 in year 3). Caution is warranted given the number of hypothesis tests in the table, but the general pattern seems to be (imprecisely) similar violence declines in year 1 but worse crime outcomes in later years for the group that replaced 2 hours per week of work with an SEL curriculum.

One potential explanation for these suggestively differential effects lies in the employment impacts shown in Table A17. The jobs + SEL group experienced significantly more crowd-out of non-program employment during the summer (a 9 percentage point decline in non-program provider employment, which is significantly different both from zero and from the jobs-only group).[11] In the remainder of year 1, the jobs+ SEL group had a significant 7 percentage point decrease in non-provider employment, which is significantly different than the employment impact in the jobs only group (p-value of difference = 0.06). The difference appears to carry over into the second post-program year, when the jobs + SEL group earns marginally less than the control group (about $248 less than controls), and about $262 less than the jobs-only group, although the difference between groups is imprecisely estimated.

If it is the case that the youth who spent more time working (rather than participating in the SEL curriculum) did somewhat better in the formal labor market, the increase in property crime among the SEL group could be partly driven by their having less money, a lower opportunity cost of time, and more free time. It is important, however, not to over-interpret these patterns for three reasons. First, Tables A15 through A17 include over 100 different hypothesis tests (including the tests of the difference across arms), so the unadjusted p-values overstate our confidence in the results. Second, our employment data are imperfect. We can only match a subset of our study youth to employment records, and the employment data do not cover the informal sector. So we cannot rule out that the jobs + SEL group developed more interpersonal skills to leverage connections to the informal labor market. If so, we could overstate the differences in employment and earnings by using only UI data. Third, the other evidence available is not always consistent with the idea that more time spent in jobs helps in the labor market while time spent in SEL has mixed effects on crime (violence decline but property crime increase). Everyone in the 2013 cohort received jobs and SEL, and we do not see significant increases in property crime; instead, we see a significant decline in year 2 drug crimes.[12] Additionally, the New York City summer jobs program, which is more like the jobs-only arm here (no SEL curriculum, though they also do not have a separate adult mentor) if anything has a small negative impact on earnings Gelber et al. (2016).

Overall, we consider the results suggestive evidence that the SEL component may not be necessary to improve youth outcomes. This is not to say, however, that the aims of the SEL curriculum are not important mechanisms. Anecdotally, employers and job mentors in the jobs-only group taught many of the same lessons about self-regulation, taking criticism, and being responsible employees as the SEL curriculum did. And the SEL was not offered in addition to the regular program; jobs + SEL youth exchanged 2 hours per day of work for SEL. So it may be that the two strategies are somewhat interchangeable. Also, as mentioned in the main text, service providers

[11]The reason for this difference is not entirely clear and could just be chance. It is possible that in the cases where the SEL curriculum could not be delivered at the youths' worksites, the extra travel time made it more difficult for youth to maintain outside employment. Or it is possible that the process of self-reflection in which SEL youth engaged led them to set priorities other than labor market involvement.

[12]This cohort is not exactly comparable, since it involves a different population of youth and the opportunity for paid post-summer activities at program providers. But it at least introduces some uncertainty about how replacing some job hours with SEL matters.

widely believed that 2 daily hours of SEL was far too much; not all of that time was spent engaged in a constructive curriculum. If youth felt like some of their time was wasted, any differences in behavior might be more attributable to that problem than to the difference between SEL and work more generally.

F.5 Subgroup Heterogeneity

In the main text, we mention that more standard approaches to estimating treatment heterogeneity - in particular, interaction effects - show few differences by subgroup. Table A18 shows estimates of the program's local average treatment effect on all our main outcomes separately by the subgroups that seem most a priori relevant to school, crime, and employment outcomes: baseline school enrollment, gender, and whether someone had at least one baseline arrest.

In general, although the magnitudes of some of the differences are suggestive, few of these effects are significantly different across subgroups. For example, the point estimate on violent-crime arrests is substantively more negative for in-school than for out-of-school youth (7.4 versus 3.5 fewer arrests per 100 participants), while the decreases in other types of arrest tend to be much larger for out-of-school youth. But none of the differences is statistically significant.

Of the 33 tests of subgroup differences in the table, 4 are significant at the 10 percent level - about what would be expected by chance. The first suggests that the program improves employment among in-school youth by connecting them with program providers but pulls out-of-school youth out of the regular labor force. The second suggests that the increase in employment may come at the cost of lower school persistence for in-school youth. The third suggests that boys have a larger increase in provider-based employment than girls do (14 versus 5 percentage points). This difference is likely driven at least in part by mechanical differences in the 2012 and 2013 programs, since all women in our sample participated in 2012, and 2013 participants were invited to participate in additional post-summer programming with the providers.

The fourth difference suggests that the violent crime decline was larger among youth who had an arrest record prior to the program (11.1 versus 2.7 fewer violent-crime arrests). Since the control complier mean for those with a baseline arrest is almost an order of magnitude larger than those with no criminal history (33.1 versus 4.2 per 100 youth), this in part reflects the fact that there is more crime to reduce for more criminally-involved youth. The proportional change is actually much larger for the youth without a prior arrest. This difference in violent crime impacts contributes to a large but not quite significant difference in the social costs crime as well (social savings of around $20,500 compared to a small decrease of about $87, p = 0.12). Youth without a prior arrest also have a larger improvement in earnings than those with a criminal history (an increase of $2,154 versus a decrease of $430, p-value of difference = 0.11).

That said, this exercise highlights why this way of testing heterogeneity in several subgroups across many outcomes is problematic. The patterns are interesting and largely logical. But even though we limited ourselves to 3 key subgroup splits, it is hard to differentiate true differences from chance findings with so many hypothesis tests. If we consider the 11 tests of the subgroup difference in a single panel as a family of outcomes, most of the significant findings become insignificant when controlling for the FDR. Only the difference which suggests male youth have a bigger increase in post-program provider employment - again, potentially a mechanical effect of the male-only post-summer programming in 2013 - remains significant (with q=0.03). This difference remains significant at the 10 percent level ($q = 0.09$) if we more conservatively group all 33 tests of subgroup differences in to a single family.

F.6 Missing Employment Data

Table A19 shows that the employment results reported in the main text are similar to the results using alternative methods to handle missing data. Recall that employment data can be missing because a youth did not have a social security number available in the CPS data (SSNs were required for matching to UI data) or was not in the CPS data at all. Our main estimates drop any observations with this kind of missing data, assigning 0s for employment or earnings only for youth who had a SSN available for matching.

Panel A of Table A19 instead makes the extreme assumption that anyone without an SSN in the data did not work, assigning a 0 for employment and wages. This is not likely to be true, but provides a sense for how much the results change with a very extreme assumption about why data are missing. Panel B instead assigns treatment or control means, calculated within randomization blocks, to all missing data. This approach assumes that data are missing completely at random (uncorrelated with observable or unobservable characteristics) after conditioning on randomization block.

Panel C relaxes this assumption to missing at random (uncorrelated only with unobservable characteristics) by using multiple imputation (MI), which takes a Bayesian approach to imputation (Little and Rubin, 2014; Puma et al., 2009). We start by regressing the outcome variable on baseline covariates, block indicators, and the non-missing outcomes (crime categories) for observations with non-missing data (separately by treatment and control groups so as not to introduce correlation between the predictions and the treatment indicator). We use the resulting parameters to predict the missing values of the outcome variable, creating an initial imputed dataset. Using the imputed dataset, we re-estimate the regression parameters, update their distributions, take new draws from the distributions, and repeat the process. After a given number of iterations, we generate a usable imputed dataset. We repeat this process 20 times, reporting coefficients that are averages across the 20 imputed data sets. The reported standard errors account for both the within- and across-imputation variances.

Although the magnitudes and statistical significance shift somewhat across imputation methods, the results are always qualitatively similar to those reported in the main text: a large increase in employment during the program with a small amount of crowd-out of non-program employment, followed by an increase in provider-based employment (but not at other employers) that does not generally translate into significant increases in overall employment or earnings. Many of the earnings point estimates are positive, but our results are imprecise enough that we generally cannot rule out the small decline in earnings seen in Gelber et al. (2016).

F.7 Missing Schooling Data

Recall that youth may be missing schooling data because they graduated, attended a charter school which did not report grade information to the district, attended a private school or a school outside of the district, or dropped out. Since these different reasons for missingness have different implications for what the missing values are likely to be, this section makes a range of different assumptions and shows that they do not substantially change the conclusions.

Since GPA can be missing even when youth have non-missing attendance data (e.g., if youth attended too few days to earn a grade or attended a charter school that does not report grades), we start with different treatments of missing GPAs. Table A20 shows alternative ways of handling missing GPA data for youth who at some point were enrolled in CPS. We exclude any youth who

graduated from CPS prior to the program start, since they do not have the potential for schooling outcomes, then test different approaches to missing GPA data for two different populations. The first row restricts attention to youth who attended at least one day of school (potentially endogenous, though we find no treatment effect on attendance), who may be more likely to have missing GPAs because their school does not report GPA information to the main data system. The second row shows results for all students with a CPS record who had not graduated prior to the program. This group likely includes more dropouts and long-term truants who are missing GPA because they are not in school.

Each column takes a different approach to imputing missing GPA values. The first column separately imputes the treatment or control means of GPA, calculated within randomization blocks, to all missing data. As mentioned in the previous subsection, this approach assumes that data are missing completely at random (uncorrelated with observable or unobservable characteristics) after conditioning on randomization block. The second column imputes this same block mean when the student attended at least 70 days of school (i.e., assumes youth should have attended enough to have grades, so are more likely to have missing data due to school reporting) and imputes zero otherwise (i.e., assumes that youth actually did not attend school and so failed to earn credits). The third column imputes block means for charter school students (charters rarely report grades) but leaves other missing observations as missing, and the fourth column imputes block means for charter school students and zero otherwise. The fifth column imputes block means for charter school students and students who attended at least 70 days of school and zero otherwise. Finally, column six uses multiple imputation using the same procedure described in Subsection F.6. Regardless of how we impute missing GPA values, we find that the program did not have a significant impact on GPAs in the first post-program school year.

Table A20 suggests that GPA results are not sensitive to how we handle missing data for youth who have a CPS record. We also show that the other schooling results are not sensitive to missing data issues. As mentioned above, many youth are missing schooling data because they were not enrolled in a CPS school, which could happen either because they transferred to non-CPS schools or because they dropped out. Table A21 shows how the main schooling results change if we treat transfers differently from dropouts. We use CPS data on verified transfers (where the central administration has confirmed that the youth transferred to a non-CPS school) to identify who is a transfer, then impute the block mean by treatment group for transfers only. As with the schooling results in the main text, which assume anyone not attending has 0 days present and that GPA is missing completely at random, we see basically no impact on enrollment, attendance, or GPA. We see a 3.4 percentage point (4%) decline in persistence through the start of the third post-program school year. When looking at this impact separately by program year, this reduction in persistence is marginally significant for the 2012 program.

In the previous tables, we have excluded youth who were not matched to a CPS record at any point in their school career (and so may have attended school outside the district for their entire lives). Table A22 shows the main schooling results when we also include these youth with imputed data. The results assume any observations without a CPS record are missing at random by using the multiple imputation procedure described in Subsection F.6. Once again, we find no significant impact on any of the schooling outcomes.[13]

[13]We note that our confidence intervals on days attended are consistent with the small positive effects seen in NYC's program. Leos-Urbel (2014) finds a 1-2 day increase in attendance, which is well within our confidence intervals.

G Benefit-Cost Comparison

This section provides a more detailed explanation of the calculations that underlie Table 6. As discussed in the main text, assigning costs to crime is an inherently uncertain exercise. This appendix outlines our approach to dealing with various sources of uncertainty and explains how we form our social cost estimates.

G.1 Source of Social Cost Estimates

The cost of crime to society comes in many parts – harm to victims (which includes direct costs like lost property or medical costs as well as indirect costs like harm and suffering, or fear and behavioral changes to avoid crime), costs to the criminal justice system (police, courts, and incarceration), and costs to the offender (lost productivity and any collateral costs of arrests and incarceration on earnings, future crime, and family).

There are two basic approaches in the literature to estimating these costs: "bottom up" and "top down." The bottom up approach focuses mostly on direct costs, combining evidence from jury awards, the costs of medical care, lost wages, and other relatively observable costs of being a victim of crime. The most widely cited estimates using this approach come from Miller, Cohen, and Wiersema (1996); we use an updated and slightly expanded version of these estimates from Cohen and Piquero (2009). The Cohen and Piquero update includes costs to the criminal justice system and approximates lost offender productivity for the small proportion of crimes that end in incarceration.

The "top down" approach includes more indirect costs like fear and behavioral changes by soliciting willingness-to-pay (WTP) for crime avoidance using contingent valuation. Conceptually, this approach may capture more of the relevant costs, but it also suffers from the typical problems of obtaining true WTP measures through survey questions. Since both top down and bottom up approaches have strengths and weaknesses, we show estimates using both versions. Our top down estimates come from Cohen and Piquero (2009)'s updated estimates of Cohen et al.'s (2004) WTP measures. We transform all dollar values into 2012 dollars using the Bureau of Labor Statistics' Consumer Price Index.

To calculate a total social cost, we assign each crime that appears in the arrest data a cost and sum all costs within an individual. To deal with the fact that arrests happen over time, we discount the costs associated with each incident based on the time of arrest relative to the end of the program (using a monthly discounting that translates into a 5% annual discount rate).

One challenge that all cost-of-crime techniques face is in assigning a statistical value of a life to fatal crimes (homicide). In practice, these costs are so large as to swamp all other crimes. This is a particular problem in finite data sets where homicide is rare, as in our data. We simply do not have the power to identify a program effect on homicide. As such, if we assigned the statistical value of a life to these incidents, we would be capitalizing on what is effectively chance in our cost estimates (whether treatment or control youth happen to have one or two more of a hugely costly outcome). To avoid this problem, we assign homicide charges the cost of an aggravated assault. This may not accurately capture the true social cost of a homicide, but it prevents our cost estimates from being dramatically swayed by an extremely rare outcome for which we lack power to estimate program effects. Topcoding the social cost of a homicide reduces the magnitude of our coefficients, so that the benefits of the program appear smaller, but also reduces the size of the (already large) standard errors by two-thirds.

G.2 Arrests versus Crimes

We measure arrests, but it is well established that only a fraction of crimes committed result in arrest (e.g., Federal Bureau of Investigation, 2014). If what we care about is the social cost of crime, we want to assign costs to all crimes, not just arrests. The common approach in the literature is to assume that crime changes in proportion to observed arrests, and multiply each arrest by an estimate of crimes-per-arrest. For example, both oft-cited Perry Preschool benefit-cost analyses take this approach (Belfield et al., 2006; Heckman et al., 2010), as do other economics of crime and cost of crime papers such as Levitt (1996) and Cohen and Piquero (2009). We use the incidence-to-arrest ratios from Cohen and Piquero (column 1 of Table 1 for arrests while under 18 and the more conservative version for adults, column 3 of Table 1).[14]

For the bottom-up estimates, we only multiply the victim costs by these scaling factors, since the costs to the criminal justice system and to offenders are only incurred when someone is actually arrested. The scaling is a little trickier for the top-down estimates, since WTP does not separate criminal justice costs from victim costs. For simplicity, we assume that people's willingness-to-pay for the criminal justice and lost offender productivity costs are the same as in the bottom-up estimates. We subtract these components of the bottom-up costs from the WTP cost estimate then take the remaining difference between top-down and bottom-up cost estimates as the victim costs and multiply that difference by the scaling factor. If people value criminal justice costs or the opportunity cost of offender time more than is reflected in the bottom-up estimates, this approach may slightly overstate the victim costs.

References

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

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal facts. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Belfield, C. R., Nores, M., Barnett, S., and Schweinhart, L. (2006). The high/scope perry preschool program cost-benefit analysis using data from the age-40 followup. *Journal of Human Resources*, 41(1):162–190.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. (1997). The

[14]It is of course possible that the program teaches youth to interact more constructively with police rather than reducing actual crime. In fact, we see a 72 percent reduction in arrests for disobeying a police officer in the first year after random assignment in our pooled sample (LATE=-2.89, SE=0.94, p<0.01, CCM=4.01). This suggests that police are not just reclassifying violent crimes as more minor crimes (which would make crimes like disobeying a police officer rise), but it may indicate that youth are learning better ways to interact in conflict situations. Even if the decline in violent crime arrests was due solely to youth more successfully avoiding detection, however, the criminal justice costs associated with an arrest and any collateral impacts from justice-system involvement would still decline.

Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study. *The Journal of Human Resources*, 32(3):549–576.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2):123–140.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, Boca Raton, FL.

Card, D., Kluve, J., and Weber, A. (2015). What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations.

Cave, G., Bos, H., Doolittle, F., and Toussaint, C. (1993). JOBSTART: Final Report on a Program for School Dropouts. Technical Report October, MDRC.

Cohen, M. A. and Piquero, A. R. (2009). New Evidence on the Monetary Value of Saving a High Risk Youth. *Journal of Quantitative Criminology*, 25:25–49.

Cohen, M. A., Rust, R. T., Steen, S., and Tidd, S. T. (2004). Willingness to pay for crime control programs. *Criminology*, 42(1):89–110.

Crépon, B. and van den Berg, G. J. (2016). Active labor market policies. *Annual Review of Economics*, 8:521–546.

Davis, J. M. and Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review: Papers and Proceedings*, 107(5):546–550.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica*, 64(5):1001–1044.

Federal Bureau of Investigation (2014). Crime in the united states. Technical report, U.S. Department of Justice.

Gelber, A., Isen, A., and Kessler, J. B. (2016). The Effects of Youth Employment: Evidence From New York City Lotteries. *The Quarterly Journal of Economics*, 131(1):423–460.

Grossman, J. B. and Sipe, C. L. (1992). Summer training and education program (step): Report on long-term impacts. Technical report, Public/Private Ventures.

Grossman, J. B., Sipe, C. L., and Millner, J. A. (1987). Summer training and education program (step): Report on 1986 experiences. Technical report, Public/Private Ventures.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Infernece, and Prediction*. Springer-Verlag, second edition.

Heckman, J. J. and Krueger, A. B. (2004). *Inequality in America*. Mit Press Cambridge, MA.

Heckman, J. J., LaLonde, R. J., and Smith, J. A. (1999). The economics and econometrics of active labor market programs. *Handbook of labor economics*, 3:1865–2097.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., and Yavitz, A. (2010). The rate of return to the highscope perry preschool program. *Journal of Public Economics*, 94(1):114–128.

Heinrich, C. J. and Holzer, H. J. (2011). Improving education and employment for disadvantaged young men: Proven and promising strategies. *The Annals of the American Academy of Political and Social Science*, 635(1):163–191.

Heller, S. B. (2014). Summer jobs reduce violence among disadvantaged youth. *Science*, 346(6214):1219–1223.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer, New York.

LaLonde, R. J. (2003). Employment and training programs. In *Means-tested Transfer Programs in the United States*, pages 517–586. University of Chicago Press.

Lehman, E. and Romano, J. P. (2005). General large sample methods. In *Testing Statistical Hypotheses*, chapter 15, pages 631–691. Springer, New York.

Leos-Urbel, J. (2014). What is a summer job worth? the impact of summer youth employment on academic outcomes. *Journal of Policy Analysis and Management*, 33(4):891–911.

Levitt, S. D. (1996). The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *The Quarterly Journal of Economics*, 111(2):319–351.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Loh, W.-Y. (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review*, 82(3):329–348.

McClanahan, W., Sipe, C., and Smith, T. (2004). Enriching summer work: An evaluation of the summer career exploration program. Technical report, Public/Private Ventures.

Mcconnell, S. and Glazerman, S. (2001). National Job Corps Study: The Benefits and Costs of Job Corps. Technical report, Mathematica Policy Research, Inc.

Millenky, M., Bloom, D., Muller-Ravett, S., and Broadus, J. (2011). Staying on Course: Three-Year Results of the National Guard Youth ChalleNGe Evaluation. Technical Report June, MDRC.

Miller, T. R., Cohen, M. A., and Wiersema, B. (1996). Victim Costs and Consequences: A New Look. Technical report, U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Morgan, J. N. (2005). History and Potential of Binary Segmentation for Exploratory Data Analysis. *Journal of Data Science*, 3:123–136.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the Analysis of Survey Data , and a Proposal. *Journal of the American Statistical Association*, 58(302):415–434.

Perez-Arce, F., Constant, L., Loughran, D. S., and Karoly, L. A. (2012). A Cost-Benefit Analysis of the National Guard Youth ChalleNGe Program. Technical report, RAND Corporation.

Puma, M. J., Olsen, R. B., Bell, S. H., and Price, C. (2009). What to do when data are missing in group randomized controlled trials. ncee 2009-0049. *National Center for Education Evaluation and Regional Assistance*.

Raphael, S. (2012). Improving Employment Prospects for Former Prison Inmates: Challenges and Policy. In Cook, P. J., Ludwig, J., and McCrary, J., editors, *Controlling Crime: Strategies and Tradeoffs*, chapter 11, pages 485–541. National Bureau of Economic Research Conference Report.

Roder, A. and Elliott, M. (2011). A promising start: Year up?s initial impacts on low-income young adults? careers. *New York: Economic Mobility Corporation*.

Romano, J. P. and Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, 138:2786–2807.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Schochet, P. Z., Burghardt, J., and McConnell, S. (2008). Does Job Corps work? Impact findings from the national Job Corps study. *American Economic Review*, 98(5):1864–1886.

Schwartz, A. E., Leos-Urbel, J., and Wiswall, M. (2015). Making Summer Matter: The Impact of Youth Employment on Academic Performance.

Stanley, M., Katz, L., and Krueger, A. (1998). Developing Skills: What We Know About The Impacts of American Employment and Training Programs on Employment, Earnings, and Educational Outcomes. Technical report, G8 Economic Summit.

Uggen, C. (2000). Work as a Turning Point in the Life Course of Criminals: A Duration Model of Age, Employment, and Recidivism. *American Sociological Review*, 65(4):529–546.

Wager, S. and Athey, S. (2015). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests.

Walker, G. and Vilella-Velez, F. (1992). Anatomy of a demonstration: The summer training and education program (step) from pilot through replication and postprogram impacts. Technical report, Public/Private Ventures.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley-Interscience.

H Tables

Table A1: Baseline Balance, Pooled Sample

	Control Mean	Control SD	Treatment Coefficient	SE	N
Demographics					
Age at Program Start	17.87	1.72	0.01	0.03	6850
Black	0.92	0.27	0.00	0.01	6850
Hispanic	0.06	0.24	0.00	0.01	6850
Arrests					
Any Baseline Arrest	0.40	0.49	0.015	0.01	6850
# Arrests: Violent	0.56	1.36	0.016	0.035	6850
# Arrests: Property	0.35	1.09	-0.01	0.028	6850
# Arrests: Drug	0.54	1.64	-0.036	0.038	6850
# Arrests: Other	1.07	2.68	-0.069	0.065	6850
Academics					
In CPS Data	0.93	0.25	0.002	0.006	6850
Engaged in CPS in June (if ever in CPS)	0.64	0.48	-0.001	0.01	6415
Prior School Year Academics if Enrolled					
Grade (if in school prior year)	10.42	1.19	-0.027	0.034	4746
Days Attended (if any attendance)	128.16	47.17	1.851	1.26	4559
Free Lunch Status (if in school prior year)	0.87	0.34	0.003	0.01	4746
GPA (if available)	2.16	0.94	-0.011	0.032	3351
Employment and Earnings					
Has SSN	0.73	0.44	0.013	0.011	6850
Worked in Prior Year (if has SSN)	0.18	0.38	-0.005	0.011	5076
Neighborhood Characteristics					
Census Tract: Median Income	34253	13657	-217.609	315.916	6850
Census Tract: Unemployment Rate	14.43	6.67	0.093	0.137	6850
Joint Significance Test	F(69,6709)=.84, p=.83				

Notes. Sample pools 2012 and 2013 cohorts. The 2012 sample includes 1634 observations, with 730 treatment and 904 control observations. The 2013 sample includes 5216 observations, with 2634 and 2582 control observations. 140 youth are in both the 2012 and 2013 samples. Balance test shows treatment coefficient and standard error clustered on individual from a regression of each characteristic on a treatment indicator, randomization block fixed effects, and duplicate application indicators. Gender not included in table since it is collinear with randomization blocks. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A2: Baseline Balance, CPS Sample

	Control Mean	Control SD	Treatment Coefficient	SE	N
Demographics					
Age at Program Start	17.82	1.74	0.003	0.026	6415
Black	0.93	0.26	0.00	0.007	6415
Hispanic	0.06	0.24	-0.003	0.006	6415
Arrests					
Any Baseline Arrest	0.43	0.49	0.014	0.01	6415
# Arrests: Violent	0.59	1.39	0.02	0.036	6415
# Arrests: Property	0.37	1.12	-0.012	0.029	6415
# Arrests: Drug	0.57	1.70	-0.049	0.04	6415
# Arrests: Other	1.14	2.75	-0.073	0.068	6415
Academics					
In CPS Data	1.00	0.00	0	0	6415
Engaged in CPS in June (if ever in CPS)	0.64	0.48	-0.001	0.01	6415
Prior School Year Academics if Enrolled					
Grade (if in school prior year)	10.42	1.19	-0.027	0.034	4746
Days Attended (if any attendance)	128.16	47.17	1.851	1.26	4559
Free Lunch Status (if in school prior year)	0.87	0.34	0.003	0.01	4746
GPA (if available)	2.16	0.94	-0.011	0.032	3351
Employment and Earnings					
Has SSN	0.79	0.41	0.012	0.01	6415
Worked in Prior Year (if has SSN)	0.18	0.38	-0.005	0.011	5076
Neighborhood Characteristics					
Census Tract: Median Income	33957	13024	-127.312	319.919	6415
Census Tract: Unemployment Rate	14.55	6.72	0.082	0.142	6415
Joint Significance Test	F(67,6274)=.8, p=.886				

Notes. Sample consists of youth who are in the Chicago Public Schools records in any year. The 2012 sample includes 1634 observations, with 730 treatment and 904 control observations. The 2013 sample includes 4781 observations, with 2437 and 2344 control observations. 140 youth are in both the 2012 and 2013 samples. Balance test shows treatment coefficient and standard error clustered on individual from a regression of each characteristic on a treatment indicator, randomization block fixed effects, and duplicate application indicators. Gender not included in table since it is collinear with randomization blocks. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A3: Baseline Balance, Employment Sample

	Control Mean	Control SD	Treatment Coefficient	SE	N
Demographics					
Age at Program Start	17.90	1.75	-0.01	0.03	5076
Black	0.93	0.26	-0.001	0.008	5076
Hispanic	0.06	0.23	-0.002	0.007	5076
Arrests					
Any Baseline Arrest	0.42	0.49	0.003	0.012	5076
# Arrests: Violent	0.60	1.42	-0.026	0.041	5076
# Arrests: Property	0.35	1.12	-0.007	0.033	5076
# Arrests: Drug	0.58	1.72	-0.06	0.045	5076
# Arrests: Other	1.13	2.81	-0.083	0.077	5076
Academics					
In CPS Data	1.00	0.00	0	0	5076
Engaged in CPS in June (if ever in CPS)	0.65	0.48	-0.017	0.011	5076
Prior School Year Academics if Enrolled					
Grade (if in school prior year)	10.48	1.19	-0.035	0.039	3714
Days Attended (if any attendance)	129.20	46.11	1.892	1.388	3567
Free Lunch Status (if in school prior year)	0.87	0.34	0.001	0.011	3714
GPA (if available)	2.18	0.93	-0.002	0.036	2669
Employment and Earnings					
Has SSN	1.00	0.00	0	0	5076
Worked in Prior Year (if has SSN)	0.18	0.38	-0.005	0.011	5076
Neighborhood Characteristics					
Census Tract: Median Income	33658	12669	-171.597	360.817	5076
Census Tract: Unemployment Rate	14.62	6.82	0.034	0.151	5076
Joint Significance Test	F(66,4958)=.81, p=.867				

Notes. Sample consists of youth who have valid social security number in the CPS data so could be matched to Unemployment Insurance records. The 2012 sample includes 1334 observations, with 603 treatment and 731 control observations. The 2013 sample includes 3742 observations, with 1913 and 1829 control observations. 117 youth are in both the 2012 and 2013 samples. Balance test shows treatment coefficient and standard error clustered on individual from a regression of each characteristic on a treatment indicator, randomization block fixed effects, and duplicate application indicators. Gender not included in table since it is collinear with randomization blocks. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A4: Baseline Balance, CPS Sample without Pre-Program Graduates

	Control Mean	Control SD	Treatment Coefficient	SE	N
Demographics					
Age at Program Start	17.39	1.70	0.02	0.03	4993
Black	0.92	0.27	0.00	0.01	4993
Hispanic	0.06	0.24	-0.01	0.01	4993
Arrests					
Any Baseline Arrest	0.45	0.50	0.02	0.01	4993
# Arrests: Violent	0.69	1.51	0.02	0.04	4993
# Arrests: Property	0.42	1.22	0.00	0.04	4993
# Arrests: Drug	0.67	1.83	-0.05	0.05	4993
# Arrests: Other	1.33	2.99	-0.08	0.08	4993
Academics					
In CPS Data	1.00	0.00	0.00	0.00	4993
Engaged in CPS in June (if ever in CPS)	0.67	0.47	0.01	0.01	4993
Prior School Year Academics if Enrolled					
Grade (if in school prior year)	10.08	0.98	-0.03	0.04	3962
Days Attended (if any attendance)	124.30	49.09	2.27	1.39	3821
Free Lunch Status (if in school prior year)	0.86	0.34	0.01	0.01	3962
GPA (if available)	2.09	0.97	-0.01	0.04	2770
Employment and Earnings					
Has SSN	0.76	0.43	0.02	0.01	4993
Worked in Prior Year (if has SSN)	0.13	0.33	-0.01	0.01	3829
Neighborhood Characteristics					
Census Tract: Median Income	33829.98	12959.70	86.33	342.57	4993
Census Tract: Unemployment Rate	14.72	6.90	0.15	0.17	4993
Joint Significance Test	F(66,4895)=.94, p=.615				

Notes. Sample consists of youth who are in the Chicago Public Schools records in any year who had not graduated prior to the program. The 2012 sample includes 1427 observations, with 644 treatment and 783 control observations. The 2013 sample includes 3566 observations, with 1866 and 1700 control observations. 97 youth are in both the 2012 and 2013 samples. Balance test shows treatment coefficient and standard error clustered on individual from a regression of each characteristic on a treatment indicator, randomization block fixed effects, and duplicate application indicators. Gender not included in table since it is collinear with randomization blocks. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A5: Multiple Hypothesis Testing Adjustments

	Program Effect		H_0: Program Effect = 0			
	CCM	LATE	Unadjusted P-value	Permuted P-value	FDR Q-value	FWER P-value
A. Arrests in Year One						
Violent	18.34	-6.38	0.00	0.01	0.04	0.03
Property	8.2	1.65	0.36	0.35	0.45	0.48
Drugs	13.89	2.3	0.43	0.48	0.45	0.48
Other	36.37	-5.06	0.27	0.27	0.45	0.48
B. Arrests in Year Two						
Violent	9.52	0.78	0.69	0.65	0.66	0.65
Property	4.21	2.95	0.12	0.11	0.24	0.28
Drugs	18.64	-5.25	0.07	0.06	0.24	0.20
Other	25.04	2.43	0.58	0.60	0.66	0.65
C. Employment in Program Quarters						
Provider Employment	0.00	1.04	0.00	0.00	0.001	0.00
Non-Provider Employment	0.16	-0.06	0.03	0.03	0.02	0.03
Earnings	122.66	1013.54	0.00	0.00	0.001	0.00
D. Employment in Post-Program Quarters						
Provider Employment	0.04	0.11	0.00	0.00	0.001	0.00
Non-Provider Employment	0.44	-0.02	0.53	0.56	0.45	0.56
Earnings	326.44	99.8	0.10	0.10	0.18	0.19
E. Schooling						
Re-enrollment	0.74	0.01	0.75	0.70	0.91	0.70
Days Present	91.39	-0.45	0.86	0.58	0.91	0.91
GPA	1.95	0.02	0.77	0.65	0.91	0.88
Persistence	0.62	-0.01	0.67	0.51	0.91	0.70

Notes. Each panel shows results for one family of outcomes. Columns 1 and 2 show control complier means and local average treatment effects (LATEs), respectively. Column 3 shows the conventional p-value of the null hypothesis that the LATE is equal to 0 from a t-distribution. Column 4 provides an alternative estimate of this p-value using the percentile of the observed t-statistic in the distribution of t-statistic estimates across 100,000 permutations of treatment status. Columns 5 and 6 show p-values which control the FWER and FDR, respectively. FDR q-values defined using unadjusted p-values in column 3.

Table A6: Participation

	Any Days	# Days		Worked Most Days	
		All	Participants	All	Participants
		A. 2012 Program			
Treatment	0.75	26.07	34.99	0.65	0.87
Control	0.00	0.00	0.00	0.00	0.00
		B. 2013 Program			
Treatment	0.30	5.45	18.05	0.09	0.29
Control	0.00	0.08	20.50	0.00	0.30
		C. 2013 Extension			
Treatment	0.20	3.69	18.53		
Control	0.00	0.18	42.45		

Notes. The 2012 sample includes 730 treatment group observations and 904 control group observations. The 2013 sample includes 2634 and 2582 treatment and control observations, respectively. "Worked Most Days" is defined as working 30 or more days in 2012 and 25 or more days in 2013.

Table A7: LATE on Number of Arrests by Year (x100), Without Baseline Covariates

Number of Arrests for:	Total	Violent	Property	Drugs	Other
		A. Pooled Sample (N=6,850)			
Year One	-8.11	-6.15***	2.03	1.91	-5.91
	(7.48)	(2.31)	(1.83)	(2.99)	(4.85)
CCM	77.44	18.11	7.83	14.28	37.22
Year Two	0.73	0.93	3.06	-4.98*	1.71
	(7.22)	(1.98)	(1.90)	(2.94)	(4.60)
CCM	57.6	9.38	4.09	18.37	25.76
All Years	-6.45	-5.35	6.30**	-3.98	-3.42
	(12.73)	(3.51)	(3.05)	(4.81)	(7.95)
CCM	144.71	29.88	12.14	35.52	67.17
		B. 2012 Sample (N=1,634)			
Year One	1.02	-3.96*	1.56	0.82	2.6
	(5.49)	(2.04)	(1.41)	(2.20)	(3.02)
CCM	26.01	9.66	3.22	3.59	9.54
Year Two	4.34	0.79	3.64**	-1.79	1.7
	(5.03)	(1.75)	(1.75)	(1.99)	(2.69)
CCM	21.95	4.18	1.51	7.49	8.77
Year Three	2.04	-0.29	2.64**	-2	1.69
	(4.94)	(1.78)	(1.33)	(2.07)	(2.73)
CCM	23.88	5.81	0.86	6.78	10.44
		C. 2013 Sample (N=5,216)			
Year One	-15.8	-7.99**	2.42	2.83	-13.07
	(13.01)	(3.88)	(3.14)	(5.19)	(8.60)
CCM	114.38	24.29	10.96	21.49	57.64
Year Two	-2.31	1.05	2.58	-7.66	1.72
	(12.61)	(3.32)	(3.14)	(5.16)	(8.17)
CCM	82.81	12.96	5.95	26.39	37.51

Notes. Regressions exclude baseline covariates other than those needed for treatment to be conditionally random (block fixed effects and duplicate application indicators). Coefficients, standard errors, and control complier means (CCMs) multiplied by 100 to show change in the number of arrests per 100 participants. "All Years" row in pooled sample includes 3 years of data for the 2012 cohort and 2 years for the 2013 cohort. Pooled sample standard errors clustered on individual; others are Huber-White. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A8: Local Average Treatment Effect on Formal Employment Outcomes, Without Baseline Covariates

Outcome:	Any Formal Employment	Any Provider Employment	Any Non-Provider Employment	All Earnings
		Panel A. Pooled Sample (N=5,076)		
Effects During Program	0.86***	1.04***	-0.06**	1022.88***
	(0.03)	(0.01)	(0.03)	(83.62)
CCM	0.12	0.00	0.16	113.32
Effects in Remaining Year One Quarters	0.03	0.04***	-0.01	67.25
	(0.03)	(0.01)	(0.03)	(172.98)
CCM	0.22	0.00	0.22	583.08
Effects in Year Two	0.03	0.09***	-0.02	170.1
	(0.03)	(0.01)	(0.03)	(262.57)
CCM	0.44	0.05	0.4	1223.19
		Panel B. 2012 Sample (N=1,334)		
Effects During Program	0.88***	1.06***	-0.08***	1259.80***
	(0.03)	(0.01)	(0.02)	(94.17)
CCM	0.1	0.00	0.17	320.3
Effects in Remaining Year One Quarters	-0.06**	-0.07**		-226.87
	(0.03)	(0.03)		(153.96)
CCM	0.22	0.22		686.68
Effects in Year Two	-0.02	0.04**	-0.04	-207.88
	(0.04)	(0.02)	(0.04)	(0.04)
CCM	0.43	0.04	0.38	0.38
		Panel C. 2013 Sample (N=3,742)		
Effects During Program	0.83***	1.02***	-0.03	809.39***
	(0.04)	(0.02)	(0.04)	(0.04)
CCM	0.14	0.00	0.15	0.15
Effects in Remaining Year One Quarters	0.11**	0.07***	0.05	332.28
	(0.05)	(0.01)	(0.05)	(0.05)
CCM	0.2	0.00	0.21	0.21
Effects in Year Two	0.07	0.13***	0	510.69
	(0.05)	(0.02)	(0.05)	(469.46)
CCM	0.44	0.04	0.42	1146.61

Notes. Regressions exclude baseline covariates other than those needed for treatment to be conditionally random (block fixed effects and duplicate application indicators). Sample includes all youth with non-missing social security numbers (N = 5,076); missing data are balanced across treatment and control groups. Any provider employment is an indicator equal to 1 if someone appeared in either program participation records or the UI data with a program agency as the employer. Any non-provider employment is an indicator equal to 1 if someone worked at an employer that did not offer the program. For 610 youth whose provider did not report earnings to the UI system, program quarter earnings equal to the wage times the number of hours reported in participation records. Negative control complier means (CCMs) set to 0. Pooled sample standard errors clustered on individual; others are Huber-White. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A9: Local Average Treatment Effect on Schooling Outcomes (Excluding Pre-Program Graduates), Without Baseline Covariates

	Any Days in Year One	# Days in Year One	GPA in Year One	Persistence through Start of Year Three
Pooled	0.012	0.264	0.017	-0.002
	(0.025)	(3.653)	(0.069)	(0.026)
CCM	0.735	90.67	1.95	0.609
N	4993	4993	2447	4993
2012	0.000	-2.829	-0.063	-0.001
	(0.016)	(3.487)	(0.070)	(0.024)
CCM	0.947	133.062	2.29	0.87
N	1427	1427	1218	1427
2013	0.025	3.389	0.188	-0.003
	(0.048)	(6.437)	(0.155)	(0.048)
CCM	0.562	56.019	1.318	0.4
N	3566	3566	1229	3566

Notes. Regressions exclude baseline covariates other than those needed for treatment to be conditionally random (block fixed effects and duplicate application indicators). Includes all youth who ever appear in the CPS data but had not graduated before the program. Attendance and grade outcomes exclude records from the schools that are part of juvenile detention and prison. GPA missing for most charter school students. Persistence equals 1 for youth who either had graduated by the end of the second post-program school year or attended at least 1 day in the third post-program school year. Pooled sample standard errors clustered on individual; others are Huber-White. CCM indicates control complier mean. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A10: Intent to Treat Program Effect on Arrests

Number of Arrests for:	Total	Violent	Property	Drugs	Other
			A. Pooled Sample (N=6,850)		
Year One	-3.02	-2.58***	0.67	0.93	-2.04
	(2.820)	(0.910)	(0.730)	(1.180)	(1.880)
CM	52.52	9.93	5.51	11.33	25.8
Year Two	0.37	0.32	1.19	-2.12*	0.98
	(2.760)	(0.790)	(0.770)	(1.170)	(1.780)
CM	48.28	7.37	5.05	12.11	23.75
All Years	-2.33	-2.34*	2.33*	-1.59	-0.73
	(4.660)	(1.360)	(1.220)	(1.860)	(3.010)
CM	106.17	18.50	11.13	24.67	51.87
			B. 2012 Sample (N=1,634)		
Year One	-0.35	-3.11**	1.24	0.45	1.07
	(3.850)	(1.520)	(1.060)	(1.660)	(2.080)
CM	25.27	6.49	3.61	4.59	10.59
Year Two	1.69	-0.1	2.85**	-1.81	0.76
	(3.610)	(1.300)	(1.320)	(1.430)	(2.020)
CM	24.11	4.90	4.41	5.63	9.18
Year Three	1.4	0.08	2.10**	-2.03	1.25
	(3.610)	(1.250)	(1.020)	(1.570)	(2.050)
CM	23.32	4.90	3.12	4.65	10.65
			C. 2013 Sample (N=5,216)		
Year One	-3.95	-2.31**	0.5	1.27	-3.4
	(3.530)	(1.100)	(0.910)	(1.470)	(2.410)
CM	72.09	11.21	7.75	17.31	35.81
Year Two	-0.43	0.39	0.62	-2.28	0.84
	(3.460)	(0.950)	(0.910)	(1.470)	(2.280)
CM	68.79	9.89	6.75	16.37	35.77

Notes. Coefficients, standard errors, and control means (CMs) multiplied by 100 to show change in the number of arrests per 100 youth offered the program. "All Years" row in pooled sample includes 3 years of data for the 2012 cohort and 2 years for the 2013 cohort. Pooled sample standard errors clustered on individual; others are Huber-White. All regressions include block fixed effects, duplicate application indicators, and the baseline covariates listed in the main text. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A11: Intent to Treat Program Effect on Formal Employment

Outcome:	Any Formal Employment	Any Provider Employment	Any Non-Provider Employment	All Earnings
		Panel A. Pooled Sample (N=5,076)		
Effects During Program	0.36***	0.44***	-0.02**	431.64***
	(0.01)	(0.01)	(0.01)	(35.35)
CM	0.23	0.02	0.21	306.85
Effects in Remaining Year One Quarters	0.01	0.02***	0.00	24.95
	(0.01)	(0.00)	(0.01)	(72.43)
CM	0.28	0.00	0.28	786.82
Effects in Year Two	0.01	0.04***	-0.01	56.1
	(0.01)	(0.01)	(0.01)	(108.95)
CM	0.45	0.03	0.43	1621.85
		Panel B. 2012 Sample (N=1,334)		
Effects During Program	0.66***	0.79***	-0.05***	919.17***
	(0.02)	(0.02)	(0.02)	(74.19)
CM	0.16	0.01	0.16	261.86
Effects in Remaining Year One Quarters	-0.04**	-0.05**	–	-150.75
	(0.02)	(0.02)		(114.09)
CM	0.19	0.18		506
Effects in Year Two	-0.01	0.03**	-0.02	-135.14
	(0.03)	(0.01)	(0.03)	(140.65)
CM	0.40	0.04	0.36	993.83
		Panel C. 2013 Sample (N=3,742)		
Effects During Program	0.25***	0.31***	-0.01	239.87***
	(0.02)	(0.01)	(0.01)	(39.76)
CM	0.25	0.03	0.23	324.82
Effects in Remaining Year One Quarters	0.03**	0.02***	0.01	96.07
	(0.02)	(0.00)	(0.02)	(90.59)
CM	0.32	0.01	0.32	899.06
Effects in Year Two	0.02	0.04***	0.00	140.81
	(0.02)	(0.01)	(0.02)	(141.63)
CM	0.47	0.03	0.46	1872.85

Notes. Sample includes all youth with non-missing social security numbers (N = 5,076); missing data are balanced across treatment and control groups. Any provider employment is an indicator equal to 1 if someone appeared in either program participation records or the UI data with a program agency as the employer. Any non-provider employment is an indicator equal to 1 if someone worked at an employer that did not offer the program. For 610 youth whose provider did not report earnings to the UI system, program quarter earnings equal to the wage times the number of hours reported in participation records. Pooled sample standard errors clustered on individual; others are Huber-White. All regressions include block fixed effects, duplicate application indicators, and the baseline covariates listed in the main text. CM indicates control mean. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A12: ITT on Schooling Outcomes, Excluding Pre-Program Graduates

	Any Days in Year One	# Days in Year One	GPA in Year One	Persistence through Start of Year Three
Pooled	0.003	-0.2	0.009	-0.004
	(0.009)	(1.157)	(0.030)	(0.010)
CM	0.69	90.204	2.062	0.578
N	4993	4993	2447	4993
2012	0	-2.337	-0.033	-0.004
	(0.010)	(2.007)	(0.040)	(0.017)
CM	0.958	136.437	2.34	0.883
N	1427	1427	1218	1427
2013	0.003	0.446	0.048	-0.006
	(0.011)	(1.391)	(0.047)	(0.012)
CM	0.566	68.91	1.774	0.438
N	3566	3566	1229	3566

Notes. Includes all youth who ever appear in the CPS data but had not graduated before the program. Attendance and grade outcomes exclude records from the schools that are part of juvenile detention and prison. GPA missing for most charter school students. Persistence equals 1 for youth who either had graduated by the end of the second post-program school year or attended at least 1 day in the third post-program school year. Pooled sample standard errors clustered on individual; others are Huber-White. All regressions include block fixed effects, duplicate application indicators, and the baseline covariates listed in the main text. CM indicates control mean. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A13: Intent to Treat Program Effect on Arrests, Poisson Regression

Crime:	Total	Violent	Property	Drugs	Other
			A. Pooled Sample		
Year One	-0.07	-0.26***	0.12	0.04	-0.09
	(0.050)	(0.090)	(0.110)	(0.090)	(0.060)
AME	-3.98	-2.58	0.81	0.57	-2.7
Year Two	-0.01	0.03	0.19	-0.16*	0.01
	(0.050)	(0.090)	(0.120)	(0.080)	(0.060)
AME	-0.45	0.22	1.18	-2.18	0.37
All Years	-0.03	-0.12*	0.18**	-0.07	-0.03
	(0.040)	(0.070)	(0.090)	(0.070)	(0.050)
AME	-4.28	-2.49	2.45	-2.06	-2.15
			B. 2012 Sample		
Year One	-0.01	-0.50**	0.32	0.11	0.13
	(0.160)	(0.220)	(0.280)	(0.360)	(0.200)
AME	-0.15	-3.26	1.16	0.51	1.37
Year Two	0.1	0.04	0.60**	-0.24	0.09
	(0.150)	(0.250)	(0.290)	(0.250)	(0.210)
AME	2.29	0.17	2.65	-1.37	0.82
Year Three	0.03	-0.09	0.58*	-0.33	0.08
	(0.150)	(0.250)	(0.300)	(0.310)	(0.180)
AME	0.60	-0.46	1.80	-1.55	0.81
			C. 2013 Sample		
Year One	-0.07	-0.22**	0.09	0.03	-0.11*
	(0.050)	(0.100)	(0.120)	(0.090)	(0.070)
AME	-5.26	-2.41	0.68	0.59	-4.05
Year Two	-0.02	0.02	0.1	-0.15*	0.01
	(0.050)	(0.100)	(0.140)	(0.090)	(0.070)
AME	-1.44	0.19	0.69	-2.43	0.19

Notes. Outcomes are arrests per 100 youth. AME indicates average marginal effects. Number of observations for pooled sample = 6850, for 2012 sample = 1634, and for 2013 sample = 5216. Crime data is through 3 (2) years post-random assignment for the 2012 (2013) cohort. Randomization block fixed effects and duplicate application indicators included in all regressions with limited set of covariates to ensure convergence. Robust standard errors (in parentheses) clustered on individual for the pooled sample. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A14: Probit Estimates of ITT on Formal Employment

Outcome:	Any Formal Employment	Any Provider Employment	Any Non-Provider Employment
	Panel A. Pooled Sample (N=5,076)		
AME During Program	0.41***	0.43***	-0.02**
	(0.02)	(0.01)	(0.01)
CM	0.23	0.02	0.21
AME in Remaining Year One Quarters	0.01	0.01	0.00
	(0.01)	(.)	(0.01)
CM	0.28	0.00	0.28
AME in Year Two	0.01	0.03***	(0.01)
	(0.02)	(0.01)	(0.02)
CM	0.45	0.03	0.43
	Panel B. 2012 Sample (N=1,334)		
AME During Program	0.87***	0.69	-0.05***
	(0.04)	(.)	(0.02)
CM	0.16	0.01	0.16
AME in Remaining Year One Quarters	-0.04**		-0.04**
	(0.02)		(0.02)
CM	0.19		0.18
AME in Year Two	-0.01	0.02**	-0.02
	(0.03)	(0.01)	(0.03)
CM	0.40	0.04	0.36
	Panel C. 2013 Sample (N=3,742)		
AME During Program	0.28***	0.31***	-0.01
	(0.02)	(0.01)	(0.01)
CM	0.25	0.03	0.23
AME in Remaining Year One Quarters	0.04**	0.01***	0.01
	(0.02)	(0.00)	(0.02)
CM	0.32	0.01	0.32
AME in Year Two	0.02	0.03***	0.00
	(0.02)	(0.01)	(0.02)
CM	0.47	0.03	0.46

Notes. Estimates are average marginal effects (AMEs) from probit regression including baseline covariates, randomization block fixed effects, and duplicate application indicators. Sample includes all youth with non-missing social security numbers (N = 5,076); missing data are balanced across treatment and control groups. Any provider employment is an indicator for any program involvement equal to 1 if someone appeared in either program participation records or the UI data with a program agency as the employer. Any non-provider employment is an indicator equal to 1 if someone worked at an employer that did not offer the program. Pooled sample standard errors clustered on individual; others are Huber-White. CM indicates control mean. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A15: ITT on Number of Arrests by Year (x100) and by Treatment Arm, 2012 Cohort

Number of Arrests for:	Total	Violent	Property	Drugs	Other
			Panel A.Year 1		
Job + Mentor	-3.4	-2.76	1.91	-0.49	-2.06
	(4.58)	(1.81)	(1.52)	(2.02)	(2.41)
Job + Mentor + SEL	2.64	-3.46**	0.59	1.37	4.14
	(5.63)	(1.74)	(1.27)	(2.58)	(2.99)
CM	23.12	7.41	2.88	3.87	8.96
P-value, test of subgroup difference	0.38	0.7	0.47	0.57	0.08
			Panel B. Year 2		
Job + Mentor	-1.35	1.03	1.03	-2.77*	-0.64
	(4.21)	(1.69)	(1.48)	(1.61)	(2.38)
Job + Mentor + SEL	4.67	-1.21	4.62**	-0.88	2.13
	(4.86)	(1.49)	(1.88)	(1.74)	(2.83)
CM	20.58	4.42	2.99	5.53	7.63
P-value, test of subgroup difference	0.28	0.23	0.09	0.28	0.40
			Panel C. Year 3		
Job + Mentor	-2.29	-0.18	1.92*	-2.12	-1.92
	(3.75)	(1.60)	(1.17)	(1.77)	(2.03)
Job + Mentor + SEL	5.01	0.33	2.27	-1.94	4.34
	(4.92)	(1.55)	(1.42)	(1.87)	(2.94)
CM	20.69	4.65	2.21	4.76	9.07
P-value, test of subgroup difference	0.14	0.79	0.83	0.92	0.04
			Panel D. Cumulative		
Job + Mentor	-7.04	-1.9	4.86*	-5.38	-4.61
	(9.32)	(3.47)	(2.66)	(3.85)	(4.74)
Job + Mentor + SEL	12.31	-4.33	7.48***	-1.45	10.60*
	(11.10)	(3.05)	(2.88)	(4.10)	(6.24)
CM	64.38	16.48	8.08	14.16	25.66
P-value, test of subgroup difference	0.14	0.51	0.46	0.41	0.04

Notes. Table shows separate intent to treat effects for the two randomly assigned treatment arms in the 2012 cohort, one of which received a social emotional learning (SEL) curriculum in place of 2 hours of daily work. Coefficients, standard errors, and control means (CMs) multiplied by 100 to show change in the number of arrests per 100 youth offered the program. Standard errors are Huber-White. All regressions include block fixed effects and the baseline covariates listed in the main text. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A16: ITT on Schooling Outcomes (Excluding Pre-Program Graduates) by Treatment Arm, 2012 Cohort

	Any Days in Year One	# Days in Year One	GPA in Year One	Persistence through Start of Year Three
Job + Mentor	0.006	-1.367	0.008	-0.009
	(0.012)	(2.346)	(0.051)	(0.021)
Job + Mentor + SEL	-0.005	-3.276	-0.074	0.000
	(0.013)	(2.607)	(0.049)	(0.022)
CM	0.958	136.437	2.34	0.883
P-value, test of subgroup difference	0.482	0.514	0.175	0.716
N	1427	1427	1218	1427

Notes. Table shows separate intent to treat effects for the two randomly assigned treatment arms in the 2012 cohort, one of which received a social emotional learning (SEL) curriculum in place of 2 hours of daily work. Includes all youth in the 2012 sample who ever appear in the CPS data but had not graduated before the program. Attendance and grade outcomes exclude records from the schools that are part of juvenile detention and prison. GPA missing for most charter school students. Persistence equals 1 for youth who either had graduated by the end of the second post-program school year or attended at least 1 day in the third post-program school year. Standard errors are Huber-White. All regressions include block fixed effects and the baseline covariates listed in the main text. CM indicates control mean. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A17: ITT on Formal Employment Outcomes by Treatment Arm, 2012 Cohort

Outcome:	Any Formal Employment	Any Provider Employment	Any Non-Provider Employment	All Earnings
		Effects During Program		
Job + Mentor	0.66***	0.79***	-0.03	993.01***
	(0.03)	(0.02)	(0.02)	(109.71)
Job + Mentor + SEL	0.65***	0.80***	-0.09***	862.22***
	(0.03)	(0.02)	(0.02)	(75.91)
CM	0.16	0.01	0.16	261.86
P-value, test of subgroup difference	0.67	0.8	0.02	0.27
		Effects in Remaining Year One Quarters		
Job + Mentor	-0.02	0.01	-0.02	4.36
	(0.03)	(0.01)	(0.02)	(169.95)
Job + Mentor + SEL	-0.07***	0.01	-0.07***	-327.08***
	(0.02)	(0.01)	(0.02)	(107.89)
CM	0.19	0	0.18	506
P-value, test of subgroup difference	0.09	0.91	0.06	0.06
		Effects in Year Two		
Job + Mentor	0.01	0.04**	0.00	-13.38
	(0.03)	(0.02)	(0.03)	(194.35)
Job + Mentor + SEL	-0.03	0.03	-0.04	-248.39*
	(0.03)	(0.02)	(0.03)	(149.72)
CM	0.4	0.05	0.36	993.83
P-value, test of subgroup difference	0.23	0.48	0.39	0.26

Notes. Table shows separate intent to treat effects for the two randomly assigned treatment arms in the 2012 cohort, one of which received a social emotional learning (SEL) curriculum in place of 2 hours of daily work. Sample includes all youth with non-missing social security numbers (N = 1,334); missing data are balanced across treatment and control groups. Any provider employment is an indicator equal to 1 if someone appeared in either program participation records or the UI data with a program agency as the employer. Any non-provider employment is an indicator equal to 1 if someone worked at an employer that did not offer the program. For 301 youth whose provider did not report earnings to the UI system, program quarter earnings equal to the wage times the number of hours reported in participation records. Standard errors are Huber-White. All regressions include block fixed effects and the baseline covariates listed in the main text. CM indicates control mean. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A18: Local Average Treatment Effects by Subgroup

	Number of Year One Arrests for:					Total Social	Any Post-Program	Any Post-Program	Any Post-Program	All	School
	Total	Violent	Property	Drugs	Other	Cost of Crime	Formal Emp.	Provider Emp.	Non-Provider Emp.	Earnings	Persistence
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
A. Treatment Heterogeneity by Baseline School Enrollment											
In School	-2.54	-7.35***	0.80	4.48	-0.48	-8830	0.07**	0.12***	0.01	218	-0.04**
	(6.35)	(2.28)	(1.80)	(2.73)	(3.93)	(6453)	(0.03)	(0.02)	(0.03)	(308)	(0.02)
CCM	59.64	16.45	7.41	8.59	27.18	41936	0.42	0.04	0.39	1504	0.87
Out of School	-21.64	-3.52	3.40	-3.86	-17.66	-9252	-0.09	0.10***	-0.10	240	0.05
	(18.54)	(5.48)	(4.47)	(7.87)	(12.89)	(14642)	(0.08)	(0.03)	(0.08)	(1107)	(0.04)
CCM	119.78	21.97	10.39	27.33	60.08	65290	0.64	0.04	0.59	2579	0.30
P-value, test of subgroup difference	0.32	0.51	0.59	0.32	0.19	0.98	0.05	0.57	0.20	0.98	0.03
N	6415	6415	6415	6415	6415	6415	5076	5076	5076	5076	6415
B. Heterogeneity by Gender											
Male	-9.71	-6.83**	1.45	3.06	-7.40	-10313	0.03	0.14***	-0.03	307	-0.02
	(9.54)	(3.00)	(2.44)	(4.01)	(6.39)	(8427)	(0.04)	(0.02)	(0.04)	(511)	(0.02)
CCM	97.35	21.94	10.38	17.98	47.04	60141	0.49	0.05	0.45	1794	0.62
Female	-1.72	-5.33**	2.30	-0.40	1.71	-5404	0.02	0.05**	0.00	11	-0.01
	(3.69)	(2.25)	(1.40)	(0.98)	(2.08)	(4906)	(0.04)	(0.02)	(0.04)	(428)	(0.02)
CCM	12.58	7.31	0.99	0.73	3.55	14160	0.45	0.01	0.43	1851	0.94
P-value, test of subgroup difference	0.44	0.69	0.76	0.40	0.18	0.61	0.81	<0.01	0.65	0.66	0.74
N	6499	6499	6499	6499	6499	6499	5076	5076	5076	5076	6415
C. Heterogeneity by Prior Arrest in ISP Data											
Has Baseline Arrest	-13.06	-11.13**	1.53	5.65	-9.11	-20535	0.01	0.10***	-0.03	-430	-0.02
	(14.27)	(4.67)	(3.61)	(5.99)	(9.78)	(12864)	(0.05)	(0.02)	(0.05)	(492)	(0.03)
CCM	137.56	33.05	14.72	23.43	66.35	87361	0.49	0.07	0.44	2154	0.53
No Baseline Arrests	-3.32	-2.66**	1.68	-0.39	-1.96	-87	0.05	0.12***	-0.01	734	-0.02
	(4.34)	(1.34)	(1.40)	(1.85)	(2.34)	(3816)	(0.04)	(0.02)	(0.04)	(556)	(0.02)
CCM	14.78	4.20	1.55	2.84	6.19	11706	0.47	0.02	0.45	1623	0.87
P-value, test of subgroup difference	0.51	0.08	0.97	0.33	0.47	0.12	0.54	0.49	0.70	0.11	0.95
N	6850	6850	6850	6850	6850	6850	5076	5076	5076	5076	6415

Notes: Table show separate local average treatment effect (LATE) estimates for select subgroups. Youth are considered in school prior to the program if they are listed as actively enrolled in the June prior to the program. LATEs are estimated using a single two stage least squares regression with indicators for treatment status and participation interacted with an indicator for being in each subgroup, including controls for being in the subgroup, block fixed effects, duplicate application indicators, and the baseline covariates listed in the main text. Panel A includes youth with a CPS record. Panels B includes youth who were not missing gender. Panel C includes all youth. Columns 7-10 include youth with non-missing UI data. Column 11 includes youth with a CPS record across all panels. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A19: LATE on Formal Employment, Missing Data Robustness Checks

Outcome:	Any Formal Employment	Any Provider Employment	Any Non-Provider Employment	All Earnings
		Panel A. Impute 0s for missing		
Effects During Program	0.69***	0.83***	-0.04**	826.27***
	(0.02)	(0.01)	(0.02)	(67.02)
CCM	0.08	-0.04	0.12	72.31
Effects in Remaining Year One Quarters	0.03	0.04***	0.00	76.65
	(0.02)	(0.01)	(0.02)	(136.54)
CCM	0.17	0.00	0.17	437.75
Effects in Year Two	0.03	0.07***	0	164.75
	(0.03)	(0.01)	(0.03)	(207.15)
CCM	0.34	0.03	0.31	937.29
		Panel B. Impute group means by block		
Effects During Program	0.89***	1.08***	-0.06***	1036.69***
	(0.02)	(0.01)	(0.02)	(66.31)
CCM	0.01	-0.18	0.17	14.58
Effects in Remaining Year One Quarters	0.04*	0.04***	0.00	88.35
	(0.02)	(0.01)	(0.02)	(135.04)
CCM	0.21	-0.01	0.21	551.9
Effects in Year Two	0.03	0.10***	-0.02	203.12
	(0.03)	(0.01)	(0.03)	(203.82)
CCM	0.42	0.03	0.4	1167.47
		Panel C. Multiple imputation		
Effects During Program	0.83***	1.05***	-0.09**	1007.38***
	(0.05)	(0.03)	(0.04)	(114.42)
CCM	0.17	-0.06	0.2	34.74
Effects in Remaining Year One Quarters	0.02	0.05***	-0.02	81.7
	(0.04)	(0.01)	(0.04)	(241.93)
CCM	0.23	-0.01	0.23	579.16
Effects in Year Two	0	0.09***	-0.05	211.18
	(0.04)	(0.03)	(0.04)	(349.43)
CCM	0.45	0.03	0.42	1184.69

Notes. Main text results exclude anyone with a missing social security number; this table uses different imputation techniques for the resulting missing employment data to include the entire sample (n = 6,850). Panel A assumes anyone not in the UI records has 0 employment and earnings, including those without SSNs. Panel B imputes treatment or control randomization block means for missing data. Panel C uses multiple imputation for missing data. Baseline covariates, randomization block fixed effects, and duplicate application indicators included in all regressions. Standard errors clustered on individual in parentheses

Table A20: LATE on GPA for Youth in CPS Data, Excluding Pre-Program Graduates, Missing Data Robustness

	Block Means	Block Means if Attended > 70 Days	Block Means if Charter Student	Block Means if Charter Student	Block Means if Charter Student or Attended >70 Days	Multiple Imputation
School Attendees	0.028	-0.002	0.021	0.024	-0.002	0.077
	(0.044)	(0.052)	(0.047)	(0.051)	(0.052)	(0.062)
CCM	1.878	1.718	1.9	1.762	1.718	1.79
N	3217	3217	3029	3217	3217	3217
Youth with CPS Record	0.035	0.009	0.037	0.028	0.009	0.013
	(0.047)	(0.048)	(0.049)	(0.049)	(0.048)	(0.058)
CCM	1.399	1.281	1.38	1.315	1.281	1.291
N	4993	4993	4805	4993	4993	4993

Notes. Sample restricted to youth with a CPS record who had not graduated prior to the program. The first row shows estimates for the sample of youth who attended at least one day of school. The second row shows estimates for all youth with a CPS record. Each column uses a different imputation of GPA. Column 1 imputes all missing GPA values with block means. Column 2 imputes missing GPA with block means for students who attended over 70 days of school and zero otherwise. Column 3 imputes missing GPA values with block means for charter school students and does not impute missing values for other students. Column 4 imputes missing GPA values with block means for charter school students and zero otherwise. Column 5 imputes missing GPA values with block means for charter school students and students who attended over 70 days of school and zero otherwise. Column 6 imputes missing GPA values using multiple imputation. All regressions estimated using two stage least squares including block fixed effects, duplicate application indicators, and the baseline covariates listed in the main text. Robust standard errors clustered on individual in parentheses. CCM indicates control complier mean. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table A21: LATE on All Schooling Outcomes for All non-CPS Graduates, Missing Data Robustness with Transfers

	Any Days in Year One	# Days in Year One	GPA in Year One	Persistence through Start of Year Three
Pooled	0.008	0.079	0.034	-0.034
	(0.022)	(2.787)	(0.050)	(0.024)
CCM	0.807	97.386	1.901	0.8
N	4993	4993	2834	4993
2012	0.006	-2.251	-0.03	-0.025*
	(0.009)	(2.356)	(0.052)	(0.014)
CCM	0.973	136.746	2.251	0.972
N	1427	1427	1252	1427
2013	0.009	1.741	0.141	-0.051
	(0.043)	(4.988)	(0.102)	(0.047)
CCM	0.676	66.005	1.375	0.673
N	3566	3566	1582	3566

Notes. The 2012 sample includes 1427 youth and the 2013 sample includes 3566 youth who have a CPS record but did not graduate before the program. Any attendance and persistence are imputed as 1 for transfers and 0 for all others with missing data. Days present and GPA are imputed with the treatment or control group means by block for transfers. Days present is imputed as 0 for non-transfers with missing data. Pooled sample standard errors clustered on individual; others are Huber-White. All regressions estimated using two stage least squares including block fixed effects, duplicate application indicators, and the baseline covariates listed in the main text. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

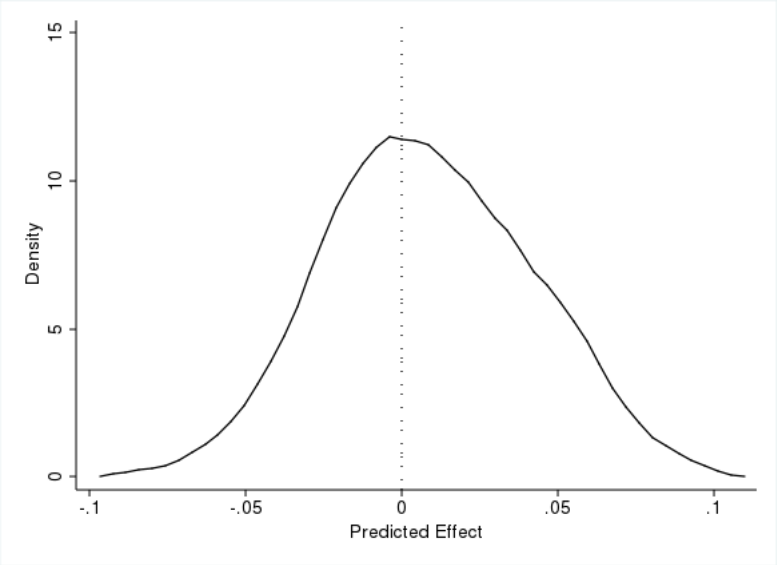
Table A22: LATE on All Schooling Outcomes for All non-CPS Graduates, Missing Data Robustness

	Any Days	# Days	GPA	Persistence Through Start of Year Three
LATE	0.005	-0.598	0.022	-0.007
	(0.019)	(2.543)	(0.054)	(0.029)
CCM	0.722	89.122	1.236	0.6

Notes. N = 5428. Main results exclude youth who never appear in CPS records; table uses multiple imputation for their outcomes. Youth who are in the CPS records but missing GPA receive an imputed 0 if they attended fewer than 70 days of school and the within-block mean of their random assignment group if they attended 70 days or over. Pre-program graduates excluded. Baseline covariates, randomization block fixed effects, and duplicate application indicators included in all regressions. Standard errors clustered on individual in parentheses. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

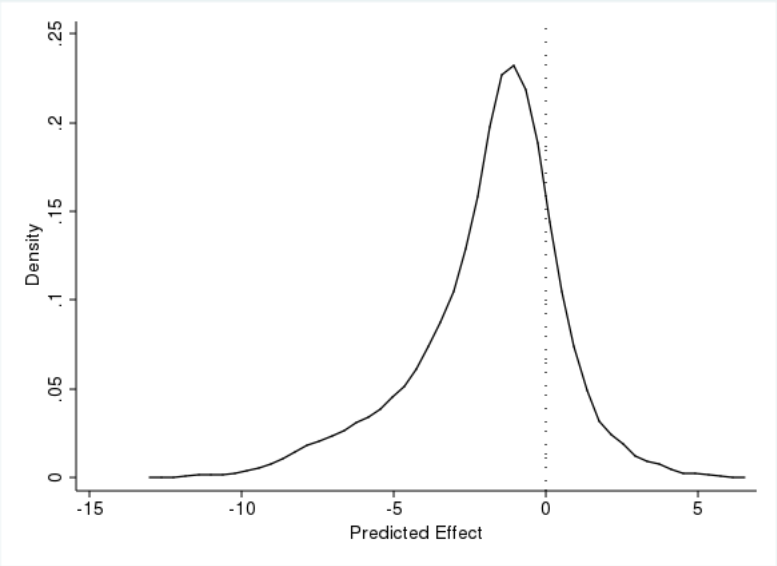
I Figures

Figure A1: Density of Predicted Employment Impacts



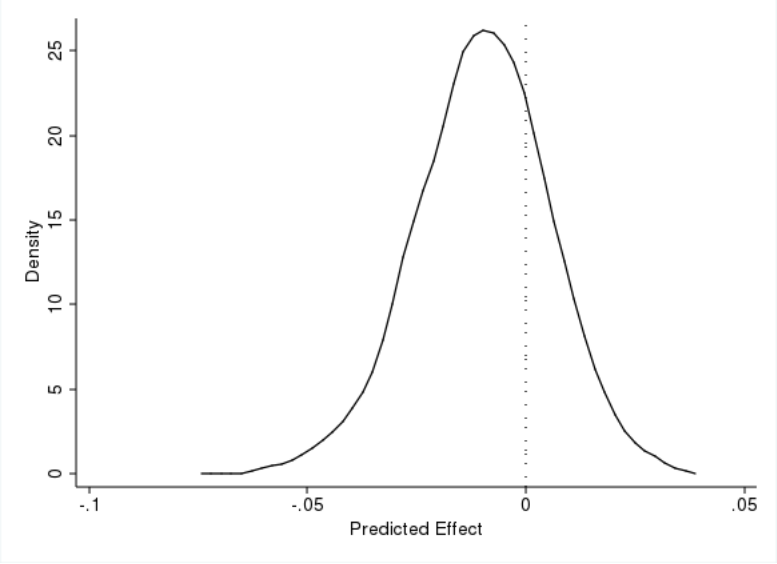
Notes. Figure shows density of predicted impacts on formal employment through 6 post-program quarters from causal forest. The average impact is 0.01, which would be a 1 percentage point increase in the probability of post-program formal employment. The dotted black line is at 0.

Figure A2: Density of Predicted Violent Crime Impacts



Notes. Figure shows density of predicted impacts on cumulative violent-crime arrests (2 years for the 2013 cohort and 3 years for the 2012 cohort) from causal forest. The average impact is -1.80 which would corresponds to 1.8 fewer arrests for violent crime per 100 youth offered the program. The dotted black line is at 0.

Figure A3: Density of Predicted School Persistence Impacts



Notes. Figure shows density of predicted impacts on school persistence through the third post-program school year from causal forest. The average impact is -.01 which would be a 1 percentage point reduction in the probability of persisting through the third post-program school year. The dotted black line is at 0.