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 Michigan and NBER

June 29, 2018

Abstract

This paper reports the results of two randomized field experiments, each offering different populations of youth a supported summer job in Chicago. The program consistently reduces violent-crime arrests, even after the summer, without improving employment, schooling, or other types of crime; if anything, property crime increases over 2-3 post-program years. Using a machine learning method to predict treatment heterogeneity, we describe who benefits and leverage the heterogeneity to explore mechanisms. We conclude that brief youth employment programs can generate substantively important behavioral change, but for different outcomes, different youth, and different reasons than those most often considered in the literature.

*Davis: University of Chicago, Saieh Hall for Economics, 1160 E 58th St, Chicago IL 60637, jonmvdavis@gmail.com. Heller: University of Michigan, 611 Tappan Ave, 238 Lorch Hall, Ann Arbor MI 48109, sbheller@umich.edu. This research was generously supported by contract B139634411 and a Scholars award from the U.S. Department of Labor, grant 2012-MIJ-FX-0002 from the Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, U.S. Department of Justice, and graduate research fellowship 2014-IJ-CX-0011 from the National Institute of Justice. The 2012 study was pre-registered at clinicaltrials.gov. Both studies are registered in the American Economic Association Registry for randomized control trials under trial numbers 1472 and 2222. For helpful comments, we thank Stephane Bonhomme, Eric Janofsky, Avi Feller, Jon Guryan, Kelly Hallberg, Jens Ludwig, Parag Pathak, Guillaume Pouliot, Sebastian Sotelo, Alexander Volfovsky, and numerous seminar participants. We are grateful to the staff of the University of Chicago Crime and Poverty Labs (especially Roseanna Ander) and the Department of Family and Support Services for supporting and facilitating the research, to Susan Athey for providing the beta causal forest code, and to Valerie Michelman and Stuart Hean for research assistance. We thank Chicago Public Schools, the Department of Family and Support Services, the Illinois Department of Employment Security, and the Illinois State Police via the Illinois Criminal Justice Information Authority for providing the data used for this analysis. The analysis and opinions here do not represent the views of any of these agencies, and any further use of the data must be approved by each agency. Any errors are our own.

1 Introduction

While adult employment has been rising in the wake of the Great Recession, youth employment over the summer - when teenagers are most likely to work - is still hovering near its 60-year low (Bureau of Labor Statistics, 2016). Young people also disproportionately suffer from violent crime. Rates of injury from violence are about twice as high among 10- to 24-year-olds than among those 25 and older, generating medical and work loss costs of over $5 billion per year (Center for Disease Control and Prevention, 2014).[1] The situation facing minority youth is even worse: African-American young people are twice as likely as their white counterparts to be unemployed, 5 times as likely to be incarcerated, and 15 times as likely to be murdered (Center for Disease Control and Prevention, 2014; Sickmund and Puzzanchera, 2014; Bureau of Labor Statistics, 2016).

For decades, policymakers have tried to address these problems by providing disadvantaged youth with a combination of job training, search assistance, remedial coursework, and subsidized work. The key idea motivating classical youth training programs, which target almost exclusively out-of-school, out-of-work youth,[2] is that providing income, improving human capital, and lowering search costs will generate better employment opportunities and reduce future reliance on public benefits (LaLonde, 2003). Improving skills and employment may in turn increase the opportunity cost of crime or improve other social outcomes, though non-employment outcomes are often treated as ancillary benefits of improved employment in the literature (Crépon and van den Berg, 2016). Reviews of the evidence on whether these programs actually achieve these goals among disadvantaged youth vary in their level of pessimism, but generally conclude that only very intensive and expensive training programs improve labor market outcomes. A tiny handful reduce crime, largely limited to the

[1] Costs are from 2010. Including fatal injuries adds another $9 billion.

[2] Almost all major large-scale employment programs targeting young people focus on these "disconnected" youth. The youth elements of the National Supported Work Demonstration, JOBSTART, the National Guard ChalleNGe, and the Job Training Partnership Act all target high school dropouts (Millenky et al., 2011; Bloom et al., 1997; Cave et al., 1993; Manpower Demonstration Research Corporation, 1980). Job Corps requires applicants to be dropouts or need additional education, training, or vocational skills, and most participants live in Job Corps centers, suggesting they are not already working and not attending school (Schochet et al., 2008). Year Up serves youth no longer in school (Roder and Elliott, 2011), and YouthBuild (a Department of Labor funded program which MDRC is in the process of evaluating) serves out-of-school and out-of-work youth.

period of the program itself (Heckman et al., 1999; Card et al., 2010; Heinrich and Holzer, 2011; King and Heinrich, 2011; Heckman and Krueger, 2004; LaLonde, 2003).[3]

Summer youth employment programs are motivated by very similar theory, but they use a somewhat different emphasis among a different population. By combining a short subsidized job with various forms of youth development, they typically aim to improve employment trajectories by focusing on basic work and soft skills among high school youth. These programs have only recently been subject to rigorous evaluation, with starkly different results from other training programs: Despite lasting only 6-8 weeks, programs in Chicago, New York, and Boston dramatically reduce violent crime and mortality, even after the program has ended (Gelber et al., 2016; Heller, 2014; Modestino, 2017). Yet evidence from New York suggests that city's program does so without improving average employment outcomes (if anything, some youth may substitute the program for private labor market activities), with small if any effects on education outcomes in NYC or Chicago (Gelber et al., 2016; Valentine et al., 2017; Leos-Urbel, 2014; Schwartz et al., 2015; Heller, 2014). This pattern of results - big crime declines after the program with no indication of improved human capital or increased opportunity costs - generates a puzzle about what summer programs are doing and why their results are so different from other youth employment programs.

Our paper uses two randomized controlled trials (RCTs) of a Chicago summer jobs program, along with a new supervised machine learning technique, to demonstrate how treatment heterogeneity helps unpack what these programs do for whom. Although some program elements varied across the two RCTs in 2012 and 2013 (see Section 2), both treatment groups were offered a 6-8 week part-time summer job at minimum wage ($8.25/hour) along with a job mentor - a constantly-available adult to assist youth in learning to be successful employees and help them deal with barriers to employment. Most youth also participated in a curriculum built on cognitive behavioral therapy principles aimed at helping them manage their cognitive and emotional responses to conflict, as well as encouraging them to set and achieve personal goals. We track youth in administrative data from the Chicago Public Schools (school records through the 2015-16 school year), the Illinois State Police (arrest records through October 2015), and the Illinois Department of Employment

[3]See Appendix A for a brief summary of the youth employment literature.

Security (Unemployment Insurance records through the first quarter of 2015).[4]

The paper makes two key additions to the literature. First, we test a successful intervention on a new population: disconnected, out-of-school youth. We purposefully recruit these youth, who look more like those typically served by training programs, to comprise about half of the second study sample. As a result, we can not only assess how well prior results generalize beyond existing study populations, but also test a candidate explanation for why summer jobs consistently reduce violent crime while other youth employment programs do not. Heller (2014) hypothesized that population differences could generate this pattern of treatment effect heterogeneity: Prevention (reaching youth before they leave school, as summer programs do) might be easier than remediation (reaching them after spells of unemployment, which is training programs' focus). But because all recently-evaluated summer programs serve almost entirely youth who are still in school,[5] it has been impossible to separate program from population differences until now.

Second, we use a new supervised machine learning technique to estimate how different youth respond to the same program. Better estimating treatment heterogeneity can help improve program targeting to generate larger social gains (Berger et al., 2001; Lechner and Smith, 2007; Frölich, 2008; Behncke et al., 2009; Bhattacharya and Dupas, 2012), as well as inform our understanding of where else the program is likely to be successful and why. Tests for heterogeneity typically involve interacting a treatment indicator with a series of baseline covariates, one at a time. But each additional hypothesis test raises the probability of spurious findings. And if heterogeneity is driven by the interaction of more than one characteristic at a time (or a non-linear function of a continuous variable), typical interaction tests may miss substantively important variation in treatment effects. To more flexibly estimate treatment heterogeneity, we use a causal forest (Wager and Athey, 2015; Athey and Imbens, 2016), predicting treatment effects based on high-dimensional, non-linear functions of observables and mining the data for responsive subgroups in a principled way.

[4]Crime results from within Chicago over the first 16 months and one-year schooling outcomes for the first 2012 RCT were reported in Heller (2014). We add a longer-term follow-up and new outcomes for that cohort (two more years of school data, two more years of crime data that now include all arrests state-wide, and previously unreported employment outcomes), as well as the entire second study in 2013.

[5]The Chicago program in Heller (2014) served all in-school youth; about 86 percent of New York's applicants are in high school with another 7 percent in college (Valentine et al., 2017); and 88 percent of the population in the Boston evaluation was still in school (Modestino, 2017).

We have previously described and assessed this method, estimating the causal forest using only half our sample to test how it performed in a hold-out sample (Davis and Heller, 2017). We demonstrated that the original causal forest procedure over-fits the data, performing well in-sample but poorly out-of-sample, and showed how to fix the problem. While the over-fitting results were clear, the halved sample size generated unstable and noisy predictions that could not answer the substantive heterogeneity questions of interest. Here we build on our prior methodological findings to avoid over-fitting, but we use all the data, doubling our sample size, as well as add 75,000 more trees to improve power and stability. These changes allow us to search for heterogeneity using all available information, describe who benefits, and use the pattern of heterogeneity to help us understand mechanisms.

We find that average treatment effects are remarkably similar across the two study populations. In both study years, a supported summer job generates dramatic and robust reductions in violent-crime arrests in the year after random assignment: the local average treatment effect is a 42 percent decline in the first study and a 33 percent decline in the second (4.2 and 7.9 fewer arrests per 100 participants, respectively). Across the whole sample, the effect is still significant after adjustments for multiple hypothesis testing. The consistency of these effects is an important result in itself given the replication crisis in the social sciences. It also provides new evidence that differences in population between summer programs and other youth employment interventions do not explain the differential crime effects across program types; summer programs reduce violent crime among disconnected youth as well.

In terms of mechanisms, the pooled sample shows a 26 percent decline in violent-crime arrests even after removing the program months from the data (p =.061), meaning the behavior change is not simply a mechanical result of keeping youth busy over the summer that disappears as soon as the job ends. The program also does not seem to increase the overall opportunity cost of crime or keep youth out of trouble more generally: Participants' total number of arrests does not change, and if anything, property crime increases in later years. Nor does it improve employment outcomes or other indicators of human capital such as schooling, at least on average.

It is possible, however, that our overall null employment effects are masking important subgroup heterogeneity that is more consistent with traditional theory about how these

programs work - one subgroup could have improved employment that drives reduced crime while another experiences crowd out that reduces employment, thereby increasing crime. In fact, the causal forest suggests that a subset of youth *do* benefit on employment. And unlike one-way interactions, the method facilitates a full description of who these youth are across all characteristics. We show that the employment benefiters are younger, more often engaged in school, more Hispanic, more female, and less likely to have an arrest record (though nearly a third of the biggest benefiters have been arrested prior to the program). In other words, the youth with the largest improvements in formal sector employment are not the disconnected youth whom other employment programs typically target.

The heterogeneity in employment, however, does not explain the violent-crime impacts. In fact, we first show that observables do not predict heterogeneity in violent-crime impacts at all; rather, it seems that everyone's violence drops.[6] To better understand mechanisms, we then use the causal forest predictions to compare other treatment impacts across youth who do versus do not show improved employment. Consistent with the idea that the program improves human capital among a subgroup, employment benefiters show a suggestive increase in school persistence, meaning they are not just substituting work for school. However, youth with no change in employment also show reduced violence involvement, while the employment benefiters actually show an increase in property-crime arrests. In other words, the pattern of employment and crime effects in the two subgroups is not consistent with the idea that employment or increased opportunity costs explain why crime changes.

So why does violence decline in the full sample, seemingly independent from changes in employment? Expanded pro-social attitudes, improved beliefs about the future, or general "staying busy" explanations are not entirely satisfactory given that property crime increases in the later follow-up years, especially among those with improved work and school outcomes. But more nuanced crime theory may help explain the results. The crime literature highlights the role of opportunity: A program that brings youth to richer areas for the first time and introduces them to new peers may increase opportunities for theft but decrease

[6]This could occur because treatment effects are homogeneous, or because effects vary by unobservables rather than observables. It is also possible that our data set is too small relative to the variation in covariates and treatment effects for the method to find significant heterogeneity for these outcomes. We note that more typical interaction effects also fail to find treatment heterogeneity that survives multiple hypothesis testing adjustments. School persistence is also not predicted by observables.

opportunities to fight, even without changing formal labor market outcomes (Cohen and Felson, 1979; Cook, 1986; Clarke, 1995). Indeed, differential impacts by crime type are fairly common in interventions that change where and with whom youth spend time (e.g., Kling et al., 2005; Deming, 2011; Jacob and Lefgren, 2003). Anecdotal evidence from employers provides another hypothesis for why violence, which by definition involves conflict with other people, may change: Employers report helping youth develop self-regulation and the ability to respond positively to criticism, which could reduce conflicts outside the workplace as well. There could also be a role for unmeasured informal sector work, peer networks, income, or violence-specific attitudes, norms, or beliefs. Further research is needed to sort out exact mechanisms. But in the meantime, we show the potential of summer jobs programs to reduce violence among a new population, as well as the potential machine learning has to help policymakers rethink who benefits from what kind of employment program and why.

2 Program Description and Experimental Design

Chicago's Department of Family and Support Services (DFSS) designed One Summer Chicago Plus (OSC+) primarily as a violence-reduction intervention. The program details varied across the two summers (discussed separately below), but the basic structure remained the same: Youth were offered a 5 hour per day, 5 day per week summer job at minimum wage ($8.25 per hour). All youth were assigned a job mentor - an adult to assist them in learning to be successful employees and to help them deal with barriers to employment - at a ratio of about 10 to 1. Characteristics of mentors varied: Some were staff at the program providers, some were college students home for the summer, and some were temporary employees from the community. Mentors participated in a one-day training (which has been revised and extended in later years of the program) and were paid a salary. DFSS administered the program through contracts with local non-profit agencies. These agencies recruited applicants, offered participating youth a brief training, hired the mentors, recruited employers, placed youth in summer jobs, provided daily lunch and bus passes when appropriate, monitored youths' progress over the course of the summer, and if youth were fired, worked with them to find an alternative placement.[7]

[7]In 2012, three agencies served as program providers: Sinai Community Institute, St. Sabina Employment Resource Center, and Phalanx Family Services. In 2013, the number of agencies grew to seven: The

One hypothesis for why prior youth employment programs require lengthy intervention to improve outcomes is that disadvantaged adolescents may lack the "soft skills" to benefit from lower-intensity programming. To test whether targeting some of these skills could improve the impact of the program, some youth also spent 2 of the 5 daily hours in a social-emotional learning (SEL) curriculum based on cognitive behavioral therapy principles.[8] This additional social-emotional training is consistent with a growing emphasis on "soft" skill curricula among summer jobs programs across the country (Ross and Kazis, 2016). The 2012 program explicitly tested the effects of replacing 2 daily hours of work with this curriculum using two treatment arms; everyone in 2013 participated in both work and SEL.

The SEL curriculum varied somewhat by provider,[9] but the lessons focused on emotional and conflict management, social information processing, and goal setting. They aimed to teach youth to understand and manage the aspects of their emotions and behavior that might interfere with successful participation and employment (e.g., the inclination, not uncommon among adolescents, to snap defensively at someone offering constructive criticism).

2.1 Summer 2012

In the first year of the program, youth ages 14 - 21 were recruited from 13 Chicago public high schools. To ensure that the study population was at risk of the key behavior of interest, the schools were chosen because they had the highest number of youth at risk of violence involvement in the city, as identified by a separate research partner. Program providers encouraged all youth attending or planning to attend these schools to apply to the program, marketing it as a summer jobs program with more work hours (and so more opportunity for income) than Chicago's standard summer programming. A total of 1,634 youth (about 13 percent of the prior year's student population in these schools) applied for the 700 available program slots.

The research team blocked youth on school and gender (the former to match youth to

Black Star Project, Blue Sky Inn, Kleo Community Family Life Center, Phalanx Family Services, St. Sabina Employment Resource Center, Westside Health Authority, and Youth Outreach Services.

[8]Over the course of the two program years, it became clear that 2 hours per day was too much time to devote to this curriculum; it has since been changed to once a week.

[9]In both years, the SEL curriculum was provided by two agencies: Youth Guidance and SGA Youth and Family Services.

the closest program provider and the latter to over-select males, who are disproportionately involved in violence). We then randomly selected 350 youth for the jobs-only treatment arm and 350 for the jobs + SEL treatment arm. Both groups had an adult job mentor. The remaining applicants were randomly ordered within blocks and treatment groups to form a waitlist. When 30 treatment youth declined to participate, the first 30 control youth (in the same block and treatment group as the decliners) were offered the program, for a total treatment group of 730.

Youth could work for a total of 8 weeks[10] at a range of employers in the non-profit and government sectors. The jobs involved tasks such as supervising younger youth at summer camps, clearing lots to plant a community garden, improving infrastructure at local schools, and providing administrative support at aldermen's offices. Because of restrictions imposed by a funder, there were no private sector jobs in this program year.

2.2 Summer 2013

In part because OSC+ was designed as an experimental program with which to test why the program works and for whom, and in part because of logistical constraints, the 2013 design differed in a few ways from the 2012 program. Because the school district lengthened the 2013-14 school year, the shorter summer necessitated a 6-week instead of an 8-week program, during which all youth received the SEL programming. Funding restrictions were lifted, so private sector jobs were included. Eligibility was limited to youth ages 16-22 in order to reduce the burden of obtaining work permits among 14- and 15-year olds. DFSS also encouraged treatment youth to keep participating in programming offered by the community service agencies after the summer ended, which included a mix of additional SEL activities, job mentoring, and social outings such as sporting events and DJ classes. These activities were much lower intensity than the summer programming, and students received a small stipend (about $200) rather than an hourly wage for participation. Because of the city's

[10]OSC+ was originally designed to run over 7 summer weeks, but additional funding allowed for an optional week-long extension of the jobs component. Eight weeks of programming were offered but not required, and in the 8th week there was no SEL programming. One service provider also offered access to additional, optional programming outside of OSC+ (like drama, graphic design, and fitness activities), but these activities were not funded by the program. Program impacts were not limited to this provider, so these activities seem unlikely to be the key driver of the results.

focus on violence reduction, DFSS also decided to limit the program to male youth.

To test for treatment heterogeneity across a broader spectrum of youth, we expanded the eligibility requirements and recruiting process to include more disconnected youth than the prior year. Participants were no longer required to be in school. The first pool of applicants (n = 2,127) was referred directly from the criminal justice system (from probation officers, juvenile detention or prison, or a center to serve justice-involved youth).[11] The rest of the applicants (n = 3,089) had applied to Chicago's broader summer programming; those who were ages 16-20, lived in one of the 30 highest-violence community areas, and included a social security number on their application[12] were entered into the lottery.

Youth were randomly assigned to treatment or control groups within applicant pool-age-geography blocks, and each block was assigned to a specific service agency. Our main analysis of the 2013 cohort consists of 5,216 youth (2,634 treatment and 2,582 control). Because of the time-constrained recruiting process, the number of youth assigned to the treatment group far exceeds the number of available slots (1,000).[13] One important implication is that the maximum take-up rate possible - even if the first thousand youth were immediately located and agreed to participate - is 38 percent (1,000 out of 2,634). Note that this is by design and should not be interpreted as indicating low demand for the program among the treatment group. Appendix C reports additional details about randomization and recruitment.

3 Data

We match our study youth to existing administrative datasets from a variety of government sources. Program application and participation records come from DFSS. We measure crime with Illinois State Police (ISP) administrative arrest records, which combine police records

[11] No one was required to apply, but adults in the justice agencies invited youth who they judged to be work-ready to fill out applications.

[12] The intention was to facilitate matching to employment records, but these hand-entered social security numbers turned out to be too error-prone for such matching. We explain our alternative source of SSNs below.

[13] In planning to serve a very mobile and arrest-prone population, it was clear that filling all the available slots would take considerable time. Rather than add to the recruiting time by giving providers the same number of names as available slots and asking them to wait for additional lists when not all youth could be located, we gave providers lists of hundreds more youth than available program slots upfront. As a result, providers were not expected to contact everyone on their lists of treatment youth. Instead, they stopped recruiting once their slots were filled. We count everyone on the treatment list as part of the treatment group, since we did not enforce the rule that providers work down the list in order.

from departments across the state.[14] Youth were probabilistically matched to these records using their name and birth date. The arrest data include the date and a description of each offense, which we use to categorize offenses as violent, property, drug, or other (vandalism, trespassing, outstanding warrants, etc.). The data cover both juvenile and adult arrests from 2001 through two (2013 cohort) or three (2012 cohort) years post-random assignment. Youth who have never been arrested will not be in the ISP records, so we assign zero arrests for individuals not matched to the data.

We use student-level administrative records from Chicago Public Schools (CPS) to capture schooling outcomes, matching youth using their unique CPS identification numbers if provided on their application, or probabilistically using their name and birth date if their numbers were unavailable. These data include details about the youths' enrollment status, grade level, course grades, and attendance[15] from the beginning of their enrollment in CPS through the 2015-16 academic year.

Missing data is of particular concern for schooling outcomes, since there are multiple reasons that youth might not appear in the data. First, 21 percent of the sample had already graduated from CPS prior to the program's start, so they could not have additional post-program high school outcomes. Second, some youth may attend private or non-Chicago public schools, which are not part of CPS records (all charter schools report attendance but many do not report grades in the administrative records). Third, some students who could be attending CPS may choose not to do so (i.e., are long-term truants or have dropped out).

Our main schooling analysis excludes pre-program graduates (n = 1,422)[16] as well as anyone who never appeared in the Chicago Public Schools records (and so likely attended

[14]Note that the prior study on the first cohort (Heller, 2014) used Chicago Police Department data rather than statewide data. Since that study only included arrests within the city of Chicago and covered a somewhat different time period, the amount of crime reported here is slightly different. For the most part, we now capture more arrests. In rare cases, we may miss some arrests that were part of the initial study, either because they have since been expunged from administrative records or because of differences in the matching process - the Illinois Criminal Justice Information Authority conducted the match to ISP data.

[15]CPS underwent a major reform of how they recorded disciplinary incidents during this time, so it is not clear how comparable recording is across or even within schools. As such, we do not use the disciplinary data as outcome measures.

[16]The lotteries actually occur about two weeks before the end of the school year, so graduating in the June before the program is not entirely a pre-program outcome. However, it seems quite unlikely that assignment to the treatment group could change graduation two weeks later. Results that only exclude those who graduated prior to the lottery are very similar.

school outside of the district for their entire lives, n = 435). Since these are both baseline characteristics, the exclusion should not undermine the integrity of random assignment (see Appendix Table A4 for balance tests on this sub-sample).[17] We focus our attendance and GPA results on the school year following the program, since missing data becomes a bigger problem over time as more students graduate, drop out, or transfer. To assess longer term academic performance, we also define a "school persistence" measure that is available for everyone in the CPS data regardless of missing attendance and GPA data in future years: an indicator that equals 1 if the youth has graduated from CPS in the first two post-program school years or is still attending school in the third post-program school year.

To measure employment, we use quarterly Unemployment Insurance (UI) records. These data include quarterly earnings, employer name, and industry for each youths' employer(s) in the formal sector. In order to match youth to UI data, the Illinois Department of Employment Security (IDES) requires youths' social security numbers (SSNs). We took advantage of the fact that the school district has historically asked for SSNs during the enrollment process.[18]

These data provide a potentially incomplete measure of employment for a number of reasons. First, as with all UI data, the records only include employment eligible for UI withholding, which excludes many agricultural and domestic positions, family employment, and any employment in the informal sector. Field work by ethnographers suggests that the informal economy may be a non-trivial source of income for youth living in low-income neighborhoods (e.g., Goffman, 2015, Venkatesh, 2006). Second, not all youth had SSNs available for matching, either because they were not in the CPS data at all (435 youth, or 6 percent of our sample), or because CPS did not have a SSN on record (1,339 of the 6,415 CPS records were missing SSNs). In all these cases, youth might have had employment records in the UI data, but they would be missing in our data. Our main analysis treats anyone without an SSN as missing, which assumes that SSNs are missing completely at random;

[17] Appendix Table A23 shows that the outcome results are similar if we impute data for students who never appear in the CPS data.

[18] Prior to May 2011, CPS asked parents and guardians to include SSNs in students' enrollment information. So study youth who were enrolled before that date had the chance to provide SSNs, although the school district did not validate them, nor require their submission. CPS provided the numbers directly to IDES without researcher involvement, and removed them before we received the data. Appendix Table A3 confirms that, since the decision to provide an SSN is a pre-program characteristic, the treatment and control groups are still balanced among the sample with non-missing data.

Appendix Table A20 shows the results are robust to different approaches to missing data.

A subset of OSC+ program providers did not report earnings to IDES. For youth attending these providers, we impute program quarter earnings as the sum of earnings at other employers and their reported program hours times \$8.25.[19] For the remaining youth with SSNs, we assign zeros for employment and earnings if youth are not in the UI data, assuming anyone not found in the matching process never worked in the formal sector. Appendix B reports additional details on all data sources and variable definitions.

4 Analytical Methods

Let Y_{ibt} denote some post-program outcome for individual i in block b during post-randomization period t. This outcome will be a function of treatment group assignment, denoted by Z_{ib}, and observed variables from administrative records measured at or before baseline, $X_{ib,t-1}$, as in equation 1 below:

$$Y_{ibt} = Z_{ib}\delta_1 + X_{ib,t-1}\delta_2 + \xi_b + u_{ibt.} \tag{1}$$

We control for the blocking variable with block fixed effects, ξ_b. The intent-to-treat effect (ITT), δ_1 in equation 1, captures the effect of being assigned to the treatment group. Although baseline characteristics are not necessary for identification, we include them in the regression to improve the precision of estimates by accounting for residual variation in the outcomes.[20]

[19] Among youth working at program providers who reported earnings, the regression coefficient of actual wages on imputed earnings is 0.81. On average, providers report earnings which are 27% higher than would be expected based on our participation records, suggesting either that our participation records understate hours worked or that program providers hired OSC+ participants for non-OSC+ opportunities at their agencies.

[20] We control for baseline covariates non-parametrically using dummy variables for categories to reduce any potential impacts of misspecification in a finite sample. Demographic controls include indicators for age at the start of the program and for being male, Black, or Hispanic. Neighborhood controls include indicators for quartiles of the census tract's unemployment rate, median income, proportion of those over 25 with a high school diploma or equivalent, and home ownership rate. Crime controls include separate indicators for having been arrested for 1 or 2 or more violent, property, drug, or other crimes. Academic controls include indicators for being in the CPS data, for having graduated prior to the program, for being enrolled in the year prior to the program (determined by June CPS enrollment status in the year of the program), for attending a neighborhood or traditional school, and for the student's free lunch status, special education status, and grade level. Our academic controls also include indicators for quartiles of number of days enrolled, for quartiles of attendance rate, and for having 1, 2, or 3 or more As, Bs, Cs, Ds, and Fs. We impute zeros for missing data and include indicator variables that equal one if a variable was missing, as well as indicator variables for submitting 1 or 2 duplicate applications. Appendix tables A7, A8, and A9 show the main crime, employment, and schooling results are substantively quite similar controlling only for

The ITT framework fully exploits the strength of the randomized experimental design. Moreover, the coefficient δ_1 in equation 1 may be useful for policy, as it directly addresses the impact of offering services on the outcome Y. But because not all youth offered the treatment participate, the ITT estimates will understate the effects of actually participating in the program on those youth who participate. Under the typical relevance and exogeneity assumptions for instrumental variables,[21] this latter set of effects can be recovered from the experimental data (Angrist et al., 1996). We perform this estimation through a two-stage least squares strategy, in which random assignment (Z_{ib}) is an instrument for program participation (P_{ibt}, an indicator variable for starting the program):

$$P_{ibt} = Z_{ib}\pi_1 + X_{ib,t-1}\pi_2 + \gamma_b + \nu_{ibt}, \tag{2}$$

$$Y_{ibt} = \hat{P}_{ibt}\beta_1 + X_{ib,t-1}\beta_2 + \alpha_b + \varepsilon_{ibt}. \tag{3}$$

If treatment effects are constant across youth, then β_1 is interpretable as the average treatment effect (ATE) across this population of disadvantaged youth, which will also equal the effects of treatment on the treated (TOT). If treatment effects are heterogeneous across youth, then β_1 represents the local average treatment effect (LATE), or the effect of treatment on youth who complied with random assignment (though in our case, with almost no control crossover, the LATE should closely approximate the TOT). To help judge the magnitude of the LATE estimates, we estimate the average outcomes of those youth in the control group who would have complied with treatment had they been assigned to treatment - the "control complier mean" (CCM) (see Heller et al. 2017, Katz et al. 2001). Because the differences between the two treatment arms in the 2012 cohort are generally not statistically significant, we focus the main text on the overall treatment-control contrast; results by treatment arm are in Appendix F.4. We report both heteroskedasticity-robust standard errors and p-values from randomization inference (permuting treatment assignment 10,000 times to approximate Fisher's exact test). The latter tests the sharp null of no treatment effects for anyone and avoids a potentially unappealing reliance on modeling assumptions and large-sample approximations that may not hold in finite samples (Athey and Imbens,

block fixed effects and having one or two duplicate applications.

[21] In order for the random assignment variable, Z_{ib}, to be a valid instrument, it must be correlated with program participation, P_{ibt}, and uncorrelated with unobservables. Moreover, if treatment effects are heterogeneous, it must shift participation in a uniform direction across people (the monotonicity assumption).

2017).

In any experiment testing program effects on multiple outcomes, not to mention heterogeneous treatment effects by subgroup, one might worry that the probability of Type I error increases with the number of tests conducted. We take a number of steps to ensure that our results are not just the result of data mining. First, we note that because DFSS built the program and recruiting strategy mainly to reduce youth violence, the impact on violent-crime arrests was the primary pre-specified outcome of interest.

Second, we present both unadjusted p-values and p-values which are adjusted using a free-step down permutation method (see Appendix D). The step-down method controls the family-wise error rate (FWER), or the probability that at least one of the true null hypotheses in a family of hypothesis tests is rejected (Anderson, 2008; Westfall and Young, 1993).[22] The FWER approach is useful for controlling the probability of making any Type I error, but it trades off power for this control. An alternative is to control the probability that a null rejection is a Type I error (the false discovery rate, or FDR), increasing the power of individual hypothesis tests in exchange for allowing some specified proportion of rejections to be false (Benjamini and Hochberg, 1995; Benjamini et al., 2006). We define our families of outcomes as: 1) the four types of crime separately for each follow-up year (violence, property, drug, and other, excluding total arrests since it is a linear combination of the rest), 2) the three main schooling outcomes across the subset of the sample that could still be in school in the post-program year (re-enrollment, days present, and GPA) plus school persistence for everyone with CPS data, 3) employment and earnings for the program quarters, and 4) employment and earnings in post-program quarters.

Third, we aim to avoid the standard approach to treatment heterogeneity: choosing

[22]We estimate the distribution of our test statistics accounting for all of the tests within a particular family by randomly permuting treatment status within blocks and recording all of the test statistics for each permutation. Under the null hypothesis of no treatment effect, each permutation should be identically distributed. Therefore, we are able to approximate the joint distribution of our test statistics with the distribution of the test statistics across permutations. For a particular hypothesis, we are able to estimate a critical value, $c(\alpha)$, with the $(1-\alpha)^{th}$ percentile of the estimated test statistic distribution. For a family of hypothesis tests, we determine the critical values using the step-down procedure outlined in Lee and Shaikh (2014). Specifically, we sort the test statistics within a family of hypothesis tests from largest to smallest. Then we determine the adjusted critical value for the test with the largest test statistic using the distribution of the maximum test statistic within the family across permutations. We then drop the test with the highest test statistic and repeat the procedure for the test with the second highest test statistic. This continues until the last test in the family. We estimate the test statistic distributions using 10,000 permutations.

several subgroups a priori to compare (or worse, testing a large number of interaction effects to find particularly responsive subgroups, which risks over-fitting and thereby detecting spurious subgroup effects). Instead, we implement a version of Wager and Athey's (2015) causal forest algorithm, which identifies who responds the most to the program by predicting treatment effects based on an individual's covariates. For this prediction, we focus on estimating conditional intent-to-treat effects, which capture differences in both youths' responses to the program and their propensity to participate if offered the program.[23] This method allows flexible, high-dimensional combinations of covariates to identify who gains from the program in a way that researcher-determined interaction effects would typically avoid.[24]

For example, suppose $\delta_{i,employment}$ is the true treatment effect on employment for an individual. Typical approaches would estimate and compare $E(\delta_{i,employment}|male = 1)$ to $E(\delta_{i,employment}|male = 0)$, or perhaps $E(\delta_{i,employment}|male = 1, African - American = 1)$ based on what the researcher specifies. If the true treatment heterogeneity is more complex than differences by gender or race (e.g., only African-American males with more than 3 prior arrests who live in neighborhoods with less than 12 percent unemployment rates benefit from the program), then researcher-specified interactions will miss it. But in theory, the causal forest can capture this pattern by searching over all values of all the covariates to isolate the combination of covariate values that predict the most heterogeneity in effects. The goal becomes predicting heterogeneity in $E(\delta_{i,employment}|X = x)$ using all the available information on Xs, rather than testing whether particular Xs are associated with significantly different treatment effects.

Our methodology for estimating causal forests, based on Athey and Imbens (2016) and

[23]The literature has not yet established that the causal forest works with IV. Take-up rates within leaves may be 0 or close to 0 because of the small samples in each leaf. This will make the LATE either incalculable or huge in the leaves resulting from some potential splits. But the causal forest implements the splits that maximize the variance of treatment effects across leaves; if some treatment effects are enormous because of small-sample variation in take-up rates, it is not clear whether the key Athey and Imbens result – that an objective function maximizing treatment effect variance is equivalent to minimizing the expected mean squared error of the unobservable prediction error – holds.

[24]There are also other supervised machine learning approaches that have this benefit, such as lasso regression and Bayesian additive regression trees (BART). The benefits and costs of each method will depend on the true form of treatment heterogeneity and its association with covariates. With the lasso, the researcher still has to specify which potential combinations of treatment interactions to include. BART constructs separate trees trained to predict Y, not the treatment effect, among treatment and controls. If different covariates predict Y than predict treatment heterogeneity, it may have difficulty predicting heterogeneity.

Wager and Athey (2015), is described in Davis and Heller (2017). We give an intuitive explanation of the steps of the method here, attempting to avoid machine learning jargon to make the discussion accessible. More complete technical details are in Appendix E. The basic goal is to divide the sample into bins that share similar covariates, and use the within-bin treatment effect as the predicted treatment effect for anyone with that bin's Xs. However, using the same observations to bin the data and predict the treatment effects within bins could induce over-fitting. So the procedure uses different subsamples for binning and for effect estimation.

To predict intent-to–treat effects conditional on covariates for a particular outcome, we repeat the following procedure: First, draw a 20 percent subsample without replacement from the data. Using a random half of the subsample, use a regression tree-based algorithm to bin the observations by values of X.[25] The algorithm recursively searches over possible ways to break the data into bins based on values of covariates, choosing the divisions that maximize the variance of treatment effects across bins subject to a penalty for within-bin variance (see appendix for algorithm details).[26] Once the bins are formed, switch to the other half of the subsample and sort the new observations into the same bins. Calculate the treatment effect ($\hat{\delta} = \bar{y}_T - \bar{y}_C$, or the difference in mean outcome between treatment and control observations) using the new observations within each bin.[27]

Next, switch to the other 80 percent of the sample (observations that are not part of the subsample), figure out in which bin each observation would belong based on its Xs, and assign that bin's $\hat{\delta}$ as the predicted treatment effect.[28] As is well established in the regression

[25] We use a subset of covariates that are available for nearly everyone in the sample including demographics (age in years and indicator variables for being male, Black, or Hispanic), neighborhood characteristics from the ACS (census tract unemployment rate, median income, proportion with at least a high school diploma, and proportion who rents their home), prior arrests (number of pre-randomization arrests for violent crime, property crime, drug crime, and other crime), prior schooling (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), and prior 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). See appendix subsection E.3 for further details.

[26] The within-bin variance penalty comes from Athey and Imbens (2016).

[27] We deal with different treatment probabilities across randomization blocks by using inverse probability weights (see appendix).

[28] This step is a slight deviation from Wager and Athey, who assign $\hat{\delta}$ to the entire sample rather than the 80 percent excluded from the initial subsample. We find that this adjustment reduces over-fitting in practice, although it may require adjusted theoretical justification (Davis and Heller, 2017).

tree literature on predicting Y instead of δ, predictions averaged across many trees have better predictive accuracy than estimates from a single tree given the high variance of a single tree's predictions (James et al., 2013). So we repeat this process with 100,000 subsamples (the causal parallel of a random forest rather than a single regression tree), averaging an observation's prediction across iterations to obtain a single predicted treatment effect. We find that increasing the number of trees from 25,000 to 100,000 dramatically increases the stability of our estimates across different random seeds.

5 Descriptive Statistics

Table 1 shows select baseline characteristics for 2012 (left panel) and 2013 (right panel) control groups, as well as tests of treatment-control balance for each covariate conditional on randomization block fixed effects. No more of the differences are significant than would be expected by chance, and tests of joint significance suggest that randomization successfully balanced the two groups (pooling both samples together, F(69,6709)=0.84 with p=0.83).

Youth in both cohorts are over 90 percent African-American and largely from poor, highly disadvantaged neighborhoods: Median neighborhood income is $33-36,000 with local unemployment rates around 13-19 percent. Thirty-eight percent of the 2012 cohort and all of the 2013 cohort is male. Recall that, in part to test for heterogeneous program effects on a broader population of youth, the eligibility rules across program years changed. As a result, the 2013 cohort is older (18.4 versus 16.3 years old), more criminally involved (47 versus 20 percent have an arrest record), and less engaged in school (51 versus 99 percent still engaged in school before the program, and those with any attendance missed 3 months versus 6 weeks of the prior school year). Partly because of their age and school status, the 2013 youth are also more likely to have been employed in the prior year (22 versus 7 percent).

6 Participation

We have two ways to measure whether a youth worked over the summer: OSC+ participation records and UI data. Participation records from the program providers are specific to OSC+, so they do not capture participation in other types of summer programs or in the regular labor force. UI data theoretically capture both, but not all study youth had SSNs to match

to the UI data, and not all program providers reported program participation to the UI system.

Our main goal is to estimate, for the complete study population, the effect of the program relative to whatever else youth would have done. As such, our first stage measures whether youth participated in OSC+ for at least one day using provider records for the entire sample. Because the nature of the counterfactual is central to understanding what this first stage is estimating, however, we also report the proportions of treatment and control youth working in other summer jobs for the subsample with available UI data.

In the first program year, 75 percent of youth offered the program actually participated, and participants averaged 35 days of work out of a possible 40. In the second program year, when the maximum possible take-up rate was 38 percent by construction (see section 2), actual program take-up was 30 percent. Participants worked an average of 18 days out of a possible 30, reflecting in part the greater challenge of recruiting and retaining a more disconnected and criminally-active population in the second year. There was no control crossover in the first cohort; 10 control youth in the second cohort (0.4 percent) participated in the program. Twenty percent of the 2013 treatment group participated in any post-summer programming. On average, these participants attended about 18.5 days of additional programming over about a 9 month period. Across both cohorts, the F-statistic on the first-stage regression measuring any OSC+ participation is 2,211.[29] See Appendix Table A6 for additional participation details.

To show what else youth were doing over the summer, Table 2 uses the sample of youth with available UI data and divides them into four mutually exclusive groups: those who worked only in OSC+ during the summer, those who worked in OSC+ and another formal sector job, those who worked only in a formal sector job, and those who did not work at all.[30] The 2012 cohort is generally less likely to be employed during the summer than the

[29]For the pooled sample, regressing a participation dummy on treatment and block fixed effects results in a coefficient of 0.4 (SE = 0.009). For the 2012 cohort, the first stage coefficient is 0.74 (SE = 0.016); for the 2013 cohort, 0.29 (SE = 0.009).

[30]UI data are quarterly, and the 2012 program started in the last week of June. So we define the "summer" program period as quarters 2 and 3 of 2012 (April - September) in the first study year and quarter 3 only (July - September) in the second study year, when the program started at the beginning of July. The table assumes anyone marked as a program participant actually worked in the program, even if they do not show up in the UI data. This can occur because some program providers considered program wages

2013 cohort: in 2012, about 8 percent of the treatment group and 15 percent of the control group work outside of OSC+, compared to 17 percent of the treatment group and 23 percent of the control group in 2013. The treatment-control differences in non-program employment suggest that OSC+ generates a small amount of crowd-out, though it still dramatically increases the overall proportion of youth who work over the summer: The treatment-control difference in having no job is 68 percentage points in 2012 (from a control mean of 84 percent) and 24 percentage points in 2013 (from a control mean of 75 percent).

7 Main Results

7.1 Crime

Table 3 shows our main crime results, which use the number of arrests of each type as the dependent variable (coefficients and standard errors are multiplied by 100, so they represent the treatment effect per 100 youth). In order to make the estimates easy to compare across program years with different take-up rates, we focus on the LATE, defining participation as any work hours greater than 0; Appendix section F.3 shows the ITT results. As described above, we have arrest data through 3 post-random assignment years for the 2012 cohort and 2 years for the 2013 cohort.

Panel A of Table 3 pools together both study cohorts, while Panels B and C show the two study cohorts separately. The patterns of behavioral change are remarkably similar across studies: both cohorts show large and statistically significant declines in violent-crime arrests during the first post-lottery year, followed in later years by declines in drug arrests but increases in property-crime arrests that vary in statistical significance. Given that the main goal of the program was violence reduction, the magnitude of the results is quite promising: the first study shows that the program causes 4.2 fewer violent-crime arrests per 100 participants in the first post-program year, a 42 percent decline. That finding is replicated in the second study, where the absolute magnitude of the change is somewhat

to be a stipend and so did not report employment to the state; the patterns of participation look almost identical when excluding the non-reporting agencies (not shown). Conversely, the table also assumes anyone who is not marked as a program participant did not participate in OSC+ (some non-participants do earn money over the summer from the same agencies that run OSC+, likely from the other summer programming those providers offer). Because not all summer programming involves UI-reported wages, we may understate broader participation in summer programming outside the formal labor market.

larger (7.9 fewer violent-crime arrests) but proportionally slightly smaller (a 33 percent decline). This pattern across cohorts is consistent with the fact that the second cohort was much more criminally active (more crime to prevent) but worked fewer hours (slightly smaller proportional change).

The findings across the two studies are both substantively similar and statistically indistinguishable across study years. This replication is important on its own, suggesting that the first study's results were not just a statistical fluke, and that a disconnected study population is unlikely to be the main reason that non-summer youth employment programs typically fail to reduce crime. Given the similarity, we focus the remainder of our discussion on Panel A, which maximizes our statistical power by pooling the study samples. In the pooled sample, the decline in violent-crime arrests during the first year is statistically significant and substantively large: 6.4 fewer violent-crime arrests per 100 participants, a 35 percent decline relative to the control complier mean. The drop is not limited to the summer of the program, when youth are mechanically kept more busy; excluding program months, violent-crime arrests decline by 26 percent (3.6 per 100 youth, $p = 0.061$). We also see positive but not statistically significant point estimates for property and drug arrests during year 1, such that - consistent with prior studies of youth employment programs that do not disaggregate crime by type - there are no significant changes in the number of total arrests.

The decline in violent-crime arrests does not continue in the second year,[31] although the initial decline is large enough that, when aggregated across all available follow-up years, the size of the cumulative violence decline is still substantively important, if somewhat imprecise (the "All Years" row shows the net effect across the entire post-random assignment period is 5.8 fewer violent-crime arrests per 100 participants relative to a control complier mean of

[31]It is worth noting that if the year 1 decline in violent-crime arrests translates into higher incarceration rates among the control youth, the year 2 results may understate the program's effects on behavior in the absence of any incarceration (incarceration temporarily reduces arrests to zero). Our estimates should be interpreted as the change in crime under the treatment regime as compared to treatment-as-usual, which includes the incapacitation effect of incarceration. Given that incarceration is socially costly - both to the government and to offenders - in theory the program could be socially beneficial, even if it has zero net effect on crime, by preventing crime at a lower cost than incarceration. We include the social costs of incarceration in our benefit-cost calculations (shown in appendix section G), so that we can ask whether spending on the program generates social benefits relative to what would have happened in the absence of the program, including the incarceration of the control group.

30.0, p = 0.082). Fade-out is almost universal in social interventions, though it is also worth noting that the program occurred at a high-violence moment in the youths' trajectories: The control complier means in year two are about half of the size of year one. This pattern suggests that part of the fade-out may stem from well-timed program delivery, after which youth start aging out of violent crime (or being incarcerated for it).

Panel A also shows a marginally significant decline in drug crimes during year 2, and imprecise but substantively large increases in property crime that, when aggregated across years, is statistically significant (5.8 more property-crime arrests per 100 participants, a 45 percent increase, p = 0.054). Program effects that go in opposite directions for violent and property crime are fairly common in the literature (e.g., Kling et al., 2005; Deming, 2011; Jacob and Lefgren, 2003); in fact, a short-term violence decline followed by a longer-term property crime increase is notably similar to the pattern of results in the Moving-to-Opportunity study. An increase in property crime might be expected if youth are spending more time traveling or working, since they have more access to better things to steal (Clarke, 1995).[32] The fact that violence is so much more socially costly than other types of crime highlights the importance of analyzing crime types separately rather than aggregating the differences away.

One obvious concern is that we are testing hypotheses across four different types of crimes over several years, and so we would expect to find a few significant effects merely due to chance. Since the division by crime type and year was determined prior to the analysis, and a decline in violence was the primary pre-specified hypothesis, one might argue that the risk of false positives generated by data mining here is quite low - especially given that it is replicated across two different studies. Nonetheless, the main finding of a year-one violence decline is robust to different adjustments for multiple hypothesis testing within years.[33] Table 6 shows the main results adjusting inference for multiple hypothesis testing

[32]It is also possible that more control compliers were incarcerated for their violent crimes during the first year of the program, which could mechanically lead to lower property crime rates among the control group during year two. However, the CCMs for drug-crime arrests are higher in year two than in year one, which is not consistent with the idea that the control youth just have less time free to offend.

[33]We perform the adjustments separately by follow-up year. This allows us to determine if the program generated any change in behavior, even in the short term. Breaking the effects down annually is useful to get a sense of the time pattern of program effects; however, we recognize that the division of effects by year is somewhat arbitrary. In the appendix, we assign social costs to outcomes and calculate the present

(see Appendix D for details on adjustments). 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 with heteroskedasticity-robust standard errors. 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 10,000 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.

Panels A and B of Table 6 show the adjustments for arrests in year one and two, respectively. The reduction in year one arrests for violent crime remains significant after adjusting the inference to control the FWER across the four crime categories (p = 0.03) or to control for the FDR (q=0.04) . The changes in other crime types over time are less robust to adjustment, and so we interpret them more cautiously. Table 6 also shows that we reject all the same null hypotheses when using randomization inference.

7.2 Schooling

One possible explanation for the violence decline could be that participants learn about the returns to schooling, or develop motivation, self-efficacy, or other pro-social beliefs, and so spend more time engaged in school in the year after the program. The schooling results in Table 4, however, suggest this is not the case: We find no significant changes in CPS re-enrollment, days present, or GPA during the school year after the program, and the confidence interval in the pooled sample rules out more than a 4-5 day increase in attendance.[34]

discounted value of the future stream of effects based on when the changes occur (testing only one program effect across the entire follow-up period).

[34]The table excludes pre-program graduates, for whom schooling data can not exist, and anyone who never appears in the CPS data, who most likely attend school outside the district. The dependent variables measure attendance in CPS, so anyone with a CPS record who does not appear in attendance records is assigned a 0 for days present. GPA is shown only for those with non-missing GPA data, which assumes that data are missing completely at random. Appendix section F.8 shows that the results are generally robust to

Table 6, Panel C shows that this remains the case after adjusting inference for multiple hypothesis testing.

The main results focus on the year after the program, since missing data becomes a larger problem as youth age (more graduation and dropout in later years). But to capture longer-term school engagement, the last column of Table 4 measures whether a youth persists in school through the start of the third year after random assignment (the dependent variable equals 1 if the youth either graduated from CPS within two post-program school years or continues to be enrolled in school through the start of the third post-program school year). The point estimate is small, negative, and statistically insignificant.

Overall, there is little evidence of changes in schooling outcomes. Longer-term analysis once more youth have had time to graduate, drop out, or attend college will be important to assess overall school effects.

7.3 Employment

Table 5 shows estimated program effects on the probability of being employed and on earnings for the sample of youth we can match to UI records.[35] As expected, there is a large increase in formal employment during the program quarters driven by greater employment at program providers.[36] There is also a small amount of crowd-out (employment outside of the program falls by about 6 percentage points), although participants' overall employment rates are about 8 times higher than among their control counterparts, leading to total summer earnings over $1,000 greater.

To exclude the mechanical program effect over the summer, we show employment during the three quarters after the program as well as the first three quarters of the second post-program year. The results in Table 5 suggest that the main effect of the program was to increase participants' attachment to employment opportunities offered by the program providers. In the 2012 cohort, not enough youth continued to work at providers during

other treatments of missing data, including logical imputation that accounts for transfers out of the district; multiple imputation, which relaxes the MCAR assumption in this sample; and the inclusion of multiply imputed data for youth who were never in CPS records.

[35] Youth are in this sample if they have a valid SSN. Appendix section F.7 shows that results using various imputation techniques for missing data do not change the pattern of results.

[36] Some coefficients are greater than 1 in part because we are using a linear probability model; Appendix Table A14 shows estimated average marginal effects using a probit, which are substantively very similar.

the post-program year to estimate a program effect, though the summer crowd-out appears to have continued past summer, when there is a decrease in employment at non-provider employers. In the 2013 cohort, the program continued into the following year, and so employment at providers increased by 7 percentage points. In both cohorts, youth appear to have formed relationships with program providers that continued in the second follow-up year, but had no significant changes in other employment or earnings. Table 6, Panels D and E show the multiple hypothesis testing adjustments for these employment outcomes. The increase in employment during the program quarters and the increase in employment at program providers after the program remain statistically significant after making these adjustments.

It is worth noting that our employment results are somewhat imprecise; for example, the top of the confidence interval on non-provider employment in year 1 for the pooled data would be a 22 percent increase. Both the crowd-out and the lack of employment increase, however, are quite similar to the findings from New York City (Gelber et al., 2016), providing some additional support for the lack of improvement (and some signs of decline) in post-summer employment outcomes.

8 Treatment Heterogeneity with the Causal Forest

We are interested in estimating treatment heterogeneity for three reasons. First, knowing who benefits most can help direct a limited resource to those with the largest potential gains.[37] Second, it may help predict external validity, since the program would seem most likely to have similar effects in other cities where youth share the characteristics of those who benefit in Chicago. If the kinds of youth driving the crime benefits are not the kinds of youth served in more classical youth training programs, heterogeneity may help explain why summer jobs have such different effects. Third, analyzing treatment heterogeneity across outcomes may help sort out the mechanisms driving the results. For example, it is possible

[37] In theory, any optimal targeting strategy should maximize net social welfare, not just behavioral benefits. Youth may generate heterogeneous program costs, if some individuals require more resources to recruit and serve, or have heterogeneous private valuations of the program. And policymakers may place value on equity or particular distributional consequences of a targeted program. As such, maximizing welfare requires taking a stand on the social welfare function, which is beyond the scope of this paper. We instead focus on estimating who benefits most, which is one crucial input to decisions about optimal allocation.

that one subgroup benefits on employment, which decreases crime, while a different subgroup experiences less employment from crowd-out and so commits more crime. But if crime benefits are not concentrated among the subgroup that benefits most from employment, then employment would seem unlikely to explain the overall violence decline.

As explained in section 4 and Appendix E.3, we use a causal forest to predict an individual's expected treatment effect for each outcome based on his covariates. Unlike more standard heterogeneity tests using interaction effects, this strategy lets us isolate the most responsive youth without limiting ourselves to single splits or linear functions of one covariate (e.g., male versus female or the linear effect of neighborhood unemployment). We focus on estimating heterogeneity in the cumulative effects of the program, using the number of violent-crime arrests over all observed post-randomization years (2 or 3 depending on the cohort), an indicator for any formal sector employment within 6 post-program quarters, and school persistence through two post-program years (still attending school or having graduated by the third post-program fall semester) as our main dependent variables.[38]

Before using the algorithm's predictions in our analysis, we first ask whether the causal forest detects any treatment heterogeneity in the data. We start with a visual presentation of how predictions and actual treatment effects are related, then quantify that performance with one simple subgroup test.[39] To see the basic pattern, we bin observations into 20 groups by percentile of predicted treatment effect. We calculate the actual ITT within each bin, then plot the predicted versus actual effects for our three main cumulative outcomes. If the predictions were perfect, we would expect the points to line up on the 45 degree line. Figures 1, 2, and 3 show these plots. The employment predictions do quite well on average, with the fitted line quite close to the 45 degree line. For the other two outcomes, it does not appear that the observables consistently predict actual variation in treatment effects.

There are many ways one could formally test the performance of the predictions, since we might be interested not just in whether the predictions linearly predict heterogeneity, but also whether they isolate any group that benefits. We choose a simple assessment of whether

[38]The overall effects on these cumulative outcomes are shown in the "all years" row of Table 3 for violence and the fourth column in Table 4 for school persistence. The LATE on overall post-program employment is 0.03, SE = 0.03, CCM = 0.48.

[39]Appendix E.4 and Figures A1-A3 show and discuss the distribution of the predictions themselves.

the group predicted to respond most positively has a significantly different treatment effect than the rest of the sample. Specifically, we create an indicator for whether a youth has a predicted treatment effect in the largest quartile of predictions (the most positive quartile for employment and school persistence and the most negative quartile for arrests). We estimate separate LATEs for this group of "predicted big responders" and the rest of the sample, and test whether the two groups have significantly different treatment effects in the data.[40]

As discussed in Davis and Heller (2017), the standard errors for these tests do not account for the fact that we are using a prediction that contains error to define our groups, because uniformly valid standard errors for the causal forest estimates have not yet been developed (and given that each prediction requires 100,000 regression trees, bootstrapping is computationally infeasible). How much to worry about the standard errors depends on what exactly we want to test. For example, prediction error might not matter if we view the groups as defined by the predictions themselves, and ask whether the groups as defined by predictions have different treatment effects; there, the group categories are an exact function of the predictions. Questions that use the predictions as a measure of some true underlying heterogeneity would be more subject to prediction error. Regardless, we try to minimize any impact of this issue by dividing the observations into two groups based on the predictions rather than using the predictions themselves; observations far from the top quartile cutoff are less likely to be misclassified. Given the imperfect standard error calculation, we emphasize the basic pattern of results rather than any single result's statistical significance. More broadly, we view the causal forest as a way to generate hypotheses to be tested in new

[40]We estimate separate effects for youth in the top quartile and bottom three quartiles using a single regression including treatment status interacted with indicators for being in each of these groups, as well as our usual baseline covariates, block fixed effects, and the main effect of being in the top quartile of predicted impacts. Appendix subsection F.5 shows the results are not sensitive to different ways of grouping by predicted effect. Davis and Heller (2017) show a similar exercise using a split-sample comparison, which produces results that are not entirely stable across different splits of the sample. Since the goal here is to learn from the predictions rather than assess the method, we make two changes to maximize power and increase stability. First, we use our full sample rather than a split sample, doubling the sample size; the substantive conclusions are similar with either method. Second, we increase the number of trees we use from 25,000 to 100,000. The predictions themselves are generally similar whether we use 25,000 or 100,000 trees (correlations across two different sets of predictions are over 0.99 for all three of our main outcomes). But since we are using a quartile cutoff to test for treatment heterogeneity, Monte Carlo error can generate small changes in predictions around the cutoff, which in turn changes the composition of our subgroups. The increase in trees reduces the Monte Carlo error, which reduces changes in quartile classification around the cutoff; the additional trees reduce the number of observations switching quartile across two different sets of predictions by 50-75 percent.

settings rather than establish ground truth (i.e., exploratory rather than confirmatory).

Table 7 shows the subgroups' LATEs across three outcomes, as well as the test of the difference. Column 1 uses an indicator for any employment over the 6 post-program quarters as the dependent variable. The group predicted to have the largest employment response has a significant 14 percentage point increase in employment, which is significantly different from the effect in the rest of the sample. In other words, the predictions successfully locate youth with large, positive treatment effects. The CCMs suggest the biggest benefiters are those who would otherwise have lower employment rates. For the other outcomes, however, the predictions are less successful. As expected, the group predicted to have the largest decline in violent-crime arrests has a more negative point estimate than the rest of the sample. But the groups are not statistically different given the large standard errors (and the difference is not entirely stable across different sets of predictions). In other words, observables do not seem to predict treatment heterogeneity - everyone in our sample benefits. Similarly, the largest quartile of predicted responders on school persistence does not have a statistically different treatment effect from the rest of the sample.[41]

As with more standard interaction tests, the failure to predict treatment heterogeneity for two outcomes could be because treatment effects are actually homogeneous or because heterogeneity is not related to observable characteristics. Consistent with (though not proof of) this possibility, Appendix E.4 shows that the causal forest predicts more variation in employment effects from the observables than it does for violence or school persistence. It is also possible that sampling variability or the form of our test obscures true variability in treatment effects, or that a larger sample with more variation in covariates would help. Based on these results, we conclude that the causal forest does identify a group who benefits from the treatment in terms of employment, which pre-specified interaction tests with adjustments

[41] We exclude pre-program graduates from the persistence column since the program could not change high school outcomes for this group. If we include these youth, the difference in persistence impacts is marginally significant across benefiters and the rest of the sample, as is the decline in persistence among non-benefiters. This may indicate that program slows down school progress for some youth, which could be consistent with the finding in Heller (2014) that youth substitute the program for summer credit recovery courses. However, the result seems to be driven by finite sample variation among pre-program graduates, whose school persistence cannot be affected by the program (adding them to the regression doubles the magnitude of the point estimate for non-benefiters). And as discussed above, we suspect our standard errors are slightly understated, making a marginally significant effect less convincing. As such, our preferred interpretation is that observables predict little clear heterogeneity in school persistence impacts.

for multiple testing would miss (see Appendix F.6).[42]

We use this variation in employment effects in two ways: to describe who benefits and to explore mechanisms. Table 8 shows pre-program descriptive statistics broken down by quartile of predicted employment treatment effects. The top row shows the mean intent-to-treat predicted effect in each quartile. The second row shows that the variation in intent-to-treat effects is not simply driven by differences in take-up rates. Although the participation rate increases a small amount across the quartiles of predicted employment effects, it is not enough to explain the differences in the intent-to-treat predictions (e.g., unadjusted for randomization block, the implied LATE for quartile 3 is 0.04 compared to 0.11 for quartile 4). The rest of the table suggests that the youth with the largest predicted employment benefits are more likely to be from the 2012 cohort (implying the positive effects are not driven solely by the low-intensity year-round programming offered only to the 2013 cohort), somewhat younger, more Hispanic, more female, and less criminally-involved than those who are not predicted to have a positive employment response (although almost a third of the biggest benefiters still have a pre-program arrest on record). The biggest responders also live in neighborhoods with somewhat lower unemployment rates, which is consistent with the possibility that labor demand plays a role in youths' ability to capitalize on their summer experience.

The employment benefiters are also more engaged in school. About 85 percent of youth in the top quartile were still in school the year before the program attending an average of 139 days of school, with 10 percent already having graduated. In the bottom quartile, by contrast, only about 46 percent of youth were still in school attending an average of 110 days of school, with 34 percent already having graduated. So on average, those who benefited most in terms of employment were more likely to be in school attending more days than those who did not show improved employment.

This descriptive exercise highlights two points. First, although the program seems to have little employment impact overall, there is a subset of participants who become more

[42]The ability to identify employment heterogeneity highlights the value-added of the causal forest relative to more standard interaction-based approaches to predicting treatment heterogeneity. Using a split-sample approach to assess the success of predictions similar to Davis and Heller (2017), a fully interacted model using the same set of covariates fails to predict significant out-of-sample treatment heterogeneity in employment (as well as in school persistence and violence).

engaged in the formal labor force. But they are not the youth whom other employment programs typically target. Most existing training programs focus on out-of-school, out-of-work youth; by contrast, the people whose employment outcomes improved in our sample - at least over the 2 post-program years in our data - tend to be younger and more engaged in school. Second, identifying the youth who benefit most is not likely to be as simple as limiting program eligibility to the characteristics that are more common among big responders, such as still being a high school student or being Hispanic. High school students are more likely to benefit, but almost half of the youth in the lowest quartile of predicted employment responders are still in school. The top quartile is more likely to be Hispanic than the bottom quartile, but nearly 85 percent of the top quartile is African-American. So simply targeting one or two characteristics may result in slightly larger gains on average, but would generally not maximize the gains from the program. Program administrators interested in maximizing program gains are therefore likely to benefit from looking at more complicated interactions of observables, or even more usefully, from a better understanding of the mechanisms that drive these differences in observable characteristics across employment responses.

Table 9 uses the causal forest predictions to explore mechanisms. The top panel shows that the employment heterogeneity is not driven by future employment at the program providers. Both employment benefiters and non-benefiters show a similar increase in future employment at program providers. The difference is in other labor market involvement, where the benefiters show an increase in non-program provider employment, while the rest of the sample shows a decrease in employment.

The remainder of the table shows how the two groups with different employment impacts respond on other outcomes. Though the standard errors are generally quite large, the table does not provide even suggestive support for the idea that employment reduces crime. The youth with no changes in employment also experience a violence decline, and property arrests go *up* among those who are working more. This pattern is not consistent with the idea that crime benefits are a result of the increased opportunity cost of crime from better employment. The results are more consistent with the idea that better employment generates more opportunities for theft, while changes in violent crime are driven by mechanisms unrelated to future employment.

The employment benefiters also have a marginally significant increase in school persistence. The statistical significance may be an artifact of our quartile cutoff (see Appendix section F.5), though the basic pattern of larger persistence point estimates for employment benefiters is consistent regardless of how we define the subgroups. The potential concentration of schooling improvements among employment benefiters suggests the additional labor force involvement is not pulling youth out of school, which is often a concern when encouraging work among school-age youth.[43] It may also be a suggestive indication that for a subset of youth, the program improves skills, motivation, or beliefs about the future, even if that does not explain why violence declines.

9 Conclusion

This paper shows that a supported summer jobs program in Chicago generates large one-year declines in violent-crime arrests, both in an initial study (42 percent decline) and in an expansion study with more disconnected youth (33 percent). The drop in violence continues after the program summer and remains substantively large after 2-3 years, though it stops accruing after the first year. And it occurs despite no detectable improvements in schooling, UI-covered employment, or other types of crime during the follow-up period. If anything, property crime increases in future years, though the large social cost of violence means that overall social benefits may still outweigh the program's administrative costs (see Appendix G). The fact that results are so similar across program years suggests that population differences can not explain why summer jobs so dramatically and reliably reduce violent crime while other youth training programs rarely do.

The standard goal of youth employment programs is to improve labor market outcomes and, as an ancillary benefit, raise the opportunity cost of all types of crime. Our pattern of results - differential effects on different types of crime with no change in employment - does not seem consistent with this mechanism. The results also seem inconsistent with other mechanisms often mentioned in relation to youth employment programs: developing pro-social beliefs or changing youths' views of their future should reduce all types of crime and perhaps improve school or employment. Providing income should reduce, not increase,

[43]Some youth are old enough that they may be balancing work with college rather than high school, which is not measured here. Future research after more youth reach college age will aim to measure this outcome.

acquisitive property crimes like theft and burglary. And keeping youth busy or out of trouble should reduce all types of crime, mostly during the program period (not reduce only violence for a year while generating later increases in property crime).

We use a new supervised machine learning method called the causal forest to test whether treatment heterogeneity is part of the explanation. In theory, some youth may have few opportunities in the absence of the program, such that OSC+ improves employment, leading to less crime. But for youth who have better work opportunities, OSC+ may crowd out a job that could turn into longer-term employment, causing an increase in crime as a result. The combination could result in zero average employment effects, with different crime effects showing up in the aggregate over time. By mining the data in a principled way, the causal forest can help us test for this kind of treatment heterogeneity.

The causal forest successfully identifies a group for whom the program improves formal sector employment. We show that this subgroup is younger and more engaged in school than the group with no employment gains - fairly different from the out-of-school and out-of-work young people usually targeted by youth employment programs. The employment benefiters also have a suggestive increase in school persistence, which may be an indication of underlying improvements in their human capital or beliefs about the future. However, the heterogeneous employment impacts do not seem to explain the pattern of crime results. The cumulative number of violent-crime arrests falls even among youth with no employment improvement. And arrests for property crime increase among those with employment gains. This is not consistent with the idea that heterogeneous changes in opportunity cost explain the heterogeneous crime effects. But it is consistent with other crime theory: Better employment provides more opportunity for theft.

A key remaining question is why violence (and little else) responds to the program in the full sample. One set of explanations involves how we are measuring outcomes. We only observe UI-covered employment, not informal employment or how youth spend their time. If unobserved time use or changes in peer networks reduce the opportunities for fighting, violence might decline with no changes in schooling or formal employment outcomes. Violence may also be better measured than other crimes. To show up in our data, a youth must be arrested after committing a crime, and clearance rates for violent crimes are considerably

higher than for other crime types (Federal Bureau of Investigation, 2016). But the property crime point estimates are in the opposite direction and sometimes statistically significant (if not entirely robust), which under-measurement seems unlikely to explain. Additionally, other interventions at the individual level often push violence in a different direction from other crimes (Kling et al., 2005; Deming, 2011; Jacob and Lefgren, 2003), and changes in labor market and macroeconomic conditions seem to affect property but not violent crime (Raphael and Winter-Ebmer, 2001; Bushway et al., 2012). So violence may just be different.

By definition, violent crime involves conflicts with other people. Learning to better avoid or manage conflict could therefore reduce violence without affecting other outcomes. Recent experimental evidence suggests that teaching self-regulation, slower decision-making, and improved social skills reduces violent crime (Heller et al., 2017; Blattman et al., 2017). And OSC+ may teach these skills: Employers and job mentors report engaging in this kind of teaching even outside the SEL curriculum. One employer reported that youths' biggest problem when they first show up to work is how defensive they are in the face of simple instructions (e.g., the need to wear close-toed shoes to work), and that he uses his time with the youth to address this tendency. Given that summer jobs programs tend to provide more of this support than what youth would otherwise receive over the summer, the additional self-regulation and conflict management skills may contribute to the reductions in violent crime.

As with any field experiment, future work will be important to assessing longer-term results and exploring program targeting, scaling, and external validity. Early labor market experience has been shown to have impacts beyond a 2-3 year follow-up period in other settings (e.g., on wage trajectories) (Holzer and LaLonde, 2000; Murphy and Welch, 1990), and some of the study youth are still in school. So longer-term follow-up is needed, especially to help assess cost effectiveness.

In terms of targeting, the main program effect is reduced violence involvement. A perhaps obvious point is that if policymakers want a program to have this effect, it would have to serve youth at some non-zero risk of violence. If policymakers preferred to focus on employment, our results suggest that targeting youth who would otherwise struggle to find work could help minimize crowd-out and generate labor force benefits, though may also

increase property crime. The causal forest suggests that younger, in-school youth may be especially likely to benefit, as are Hispanic youth and those living in areas with slightly lower (though given our study population, still high) unemployment rates.

That said, using a summer jobs program to increase youth employment at a large scale involves its own challenges. Maintaining program quality at scale is a consistent policy challenge. And expanding the number of program slots might lead employers to offer fewer jobs to non-program youth. This could shift who obtains jobs without generating an overall improvement in employment (as in Crépon et al., 2013). If the goal of increased scale is to improve employment among disadvantaged youth, the distributional consequences of scale - who, if anyone, ends up displaced - may matter.

There tends to be a fair amount of pessimism in the youth employment literature about how difficult and costly it is to improve youth outcomes. The evidence we present here, combined with growing evidence from programs in other cities, suggests that this pessimism may stem in part from mistaken beliefs about what these programs achieve and for whom. The consensus in the literature is that only long and expensive interventions can improve human capital in a way that has a lasting employment effect among the most disconnected youth. But it may take less investment to generate a large change in a very socially costly outcome like violence, or even to improve employment among a younger group of prepared students who would otherwise struggle to find work.

References

Anderson, Michael L., "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*, 2008, *103* (484), 1481–1495.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 1996, *91* (434), 444–455.

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

_ **and Guido W. Imbens**, "The Econometrics of Randomized Experiments," in Abhijit Vinayak Banerjee and Esther Duflo, eds., *Handbook of Field Experiments, Vol. 1*, 2017, chapter 3, pp. 73–140.

Behncke, Stefanie, Markus Frölich, and Michael Lechner, "Targeting Labour Market Programmes - Results from a Randomized Experiment," *Swiss Journal of Economics and Statistics*, 2009, (3), 221–268.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, 2006, *93* (3), 491–507.

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

Berger, Mark C., Dan Black, and Jeffrey A. Smith, "Evaluating profiling as a means of allocating government services," in Michael Lechner and Friedhelm Pfeiffer, eds., *Econometric Evaluation of Labour Market Policies*, Heidelberg: Physica-Verlag HD, 2001, pp. 59–84.

Bhattacharya, Debopam and Pascaline Dupas, "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 2012, pp. 168–196.

Blattman, Christopher, Julian C. Jamison, and Margaret Sheridan, "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia," *American Economic Review*, 2017, *107* (4), 1165–1206.

Bloom, Howard S, Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos, "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *The Journal of Human Resources*, 1997, *32* (3), 549–576.

Bureau of Labor Statistics, "Employment and Unemployment Among Youth - Summer 2016," Technical Report August 2016.

Bushway, Shawn, Phillip J. Cook, and Matthew Phillips, "The Overall Effect of the Business Cycle on Crime," *German Economic Review*, 2012, *13* (4), 436–446.

Card, David, Jochen Kluve, and Andrea Weber, "Active labour market policy evaluations: a meta-analysis*," *The Economic Journal*, 2010, *120* (548), F452–F477.

Cave, George, Hans Bos, Fred Doolittle, and Cyril Toussaint, "JOBSTART: Final Report on a Program for School Dropouts," Technical Report October, MDRC 1993.

Center for Disease Control and Prevention, "Web-based Injury Statistics Query and Reporting System (WISQARS)," 2014.

Clarke, Ronald V., "Situational Crime Prevention," *Crime and Justice*, 1995, *19 ,Building a Safer Society: Strategic Approaches to Crime Prevention*, 91–150.

Cohen, Lawrence E. and Marcus Felson, "Social Change and Crime Rate Trends: A Routine Activity Approach," *American Sociological Review*, 1979, *44* (4), 588–608.

Cook, Phillip J., "The Demand and Supply of Criminal Opportunities," *Crime and Justice*, 1986, *7*, 1–27.

Crépon, Bruno and Gerard J. van den Berg, "Active Labor Market Policies," *Annual Review of Economics*, 2016, *8*, 521–546.

_ , Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment," *Quarterly Journal of Economics*, 2013, *128* (2), 531–580.

Davis, Jonathan M.V. and Sara B. Heller, "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs," *American Economic Review: Papers and Proceedings*, 2017, *107* (5), 546–550.

Deming, David J, "Better schools, less crime," *The Quarterly Journal of Economics*, 2011, *126* (4), 2063–2115.

Federal Bureau of Investigation, "Crime in the United States, 2015," Technical Report September, United States Department of Justice 2016. Date accessed: 13/03/2017.

Frölich, Markus, "Statistical Treatment Choice: An Application to Active Labor Market Programs," *Journal of the American Statistical Association*, 2008, *103* (428), 547–558.

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

Goffman, Alice, *On The Run: Fugitive Life in an American City*, Macmillan, 2015.

Heckman, James J and Alan B Krueger, *Inequality in America*, Mit Press Cambridge, MA, 2004.

_ , Robert J LaLonde, and Jeffrey A Smith, "The economics and econometrics of active labor market programs," *Handbook of labor economics*, 1999, *3*, 1865–2097.

Heinrich, Carolyn J and Harry J Holzer, "Improving education and employment for

disadvantaged young men: Proven and promising strategies," *The Annals of the American Academy of Political and Social Science*, 2011, *635* (1), 163–191.

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

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack, "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago," *Quarterly Journal of Economics*, 2017, *132* (1), 1–54.

Holzer, Harry J. and Robert J. LaLonde, "Job Change and Job Stability Among Less Skilled Young Workers," in David E. Card and Rebecca M. Blank, eds., *Finding Jobs: Work and Welfare Reform*, Russell Sage Foundation, 2000.

Jacob, Brian and Lars Lefgren, "Are Idle Hands the Devil's Workshop? Incapacitation, Concentration and Juvenile Crime," *American Economic Review*, 2003, *93* (5), 1560–1577.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning*, Vol. 7, New York: Springer, 2013.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman, "Moving to opportunity in boston: early results of a randomized mobility experiment* l," *The Quarterly Journal of Economics*, 2001, *116* (May), 607–654.

King, Christopher T and Carolyn J Heinrich, "How Effective Are Workforce Development Programs? Implications for U.S. Workforce Policies," Technical Report November 2011.

Kling, Jeffrey R, Jens Ludwig, and Lawrence F Katz, "Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment," *The Quarterly Journal of Economics*, 2005, pp. 87–130.

LaLonde, Robert J, "Employment and training programs," in "Means-tested Transfer Programs in the United States," University of Chicago Press, 2003, pp. 517–586.

Lechner, Michael and Jeffrey Smith, "What is the value added by caseworkers?," *Labour Economics*, 2007, *14*, 135–151.

Lee, Soohyung and Azeem M. Shaikh, "Multiple Testing and Heterogenous Treatment Effects: Re-Evaluating the Effect of Progresa on School Enrollment," *Journal of Applied Econometrics*, 2014, *29*, 612–626.

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

Manpower Demonstration Research Corporation, *Summary and Findings of the National Supported Work Demonstration*, Cambridge, MA: Ballinger Publishing Company,

1980.

Millenky, Megan, Dan Bloom, Sara Muller-Ravett, and Joseph Broadus, "Staying on Course: Three-Year Results of the National Guard Youth ChalleNGe Evaluation," Technical Report June, MDRC 2011.

Modestino, Alicia Sasser, "How Do Summer Youth Employment Programs Improve Criminal Justice Outcomes, and for Whom?," *Federal Reserve Bank of Boston Community Development Discussion Paper*, 2017, *2017-01.*

Murphy, Kevin M. and Finis Welch, "Empirical Age-Earnings Profiles," *Journal of Labor Economics*, 1990, *8* (2), 202–229.

Raphael, Steven and Rudolf Winter-Ebmer, "Identifying the Effect of Unemployment on Crime," *The Journal of Law and Economics*, 2001, *44* (1), 259–283.

Roder, Anne and Mark Elliott, "A Promising Start: Year Up?s Initial Impacts on Low-Income Young Adults? Careers," *New York: Economic Mobility Corporation*, 2011.

Ross, Martha and Richard Kazis, "Youth Summer Jobs Programs: Aligning Ends and Means," *Metropolitan Policy Program at Brookings*, 2016.

Schochet, Peter Z., John Burghardt, and Sheena McConnell, "Does Job Corps work? Impact findings from the national Job Corps study," *American Economic Review*, 2008, *98* (5), 1864–1886.

Schwartz, Amy Ellen, Jacob Leos-Urbel, and Matthew Wiswall, "Making Summer Matter: The Impact of Youth Employment on Academic Performance," *NBER Working Paper*, 2015, *21470.*

Sickmund, Melissa and Charles Puzzanchera, "Juvenile offenders and victims: 2014 national report," Technical Report, National Center for Juvenile Justice 2014.

Valentine, Erin Jacobs, Chloe Anderson, Farhana Hossain, and Rebecca Unterman, "An Introduction to the World of Work: A Study of the Implementation and Impacts of New York City's Summer Youth Employment Program," *MDRC*, 2017.

Venkatesh, Sudhir Alladi, *Off the Books: The Underground Economy of the Urban Poor*, Harvard University Press, 2006.

Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," 2015.

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

10 Tables and Figures

Table 1: Baseline Balance

Program Year:	2012					2013				
	N	Control Mean	Control SD	Treatment Coefficient	SE	N	Control Mean	Control SD	Treatment Coefficient	SE
Demographics										
Age at Program Start	1,634	16.30	1.45	-0.05	(0.069)	5,216	18.42	1.45	0.03	(0.024)
Black	1,634	0.96	0.21	0.00	(0.010)	5,216	0.91	0.29	0.01	(0.009)
Hispanic	1,634	0.03	0.17	0.00	(0.007)	5,216	0.07	0.25	0.00	(0.008)
Arrests										
Any Baseline Arrest	1,634	0.20	0.40	0.01	(0.020)	5,216	0.47	0.50	0.02	(0.012)
# Arrests: Violent	1,634	0.15	0.56	0.03	(0.030)	5,216	0.70	1.52	0.01	(0.045)
# Arrests: Property	1,634	0.10	0.44	-0.01	(0.021)	5,216	0.44	1.23	-0.01	(0.036)
# Arrests: Drug	1,634	0.06	0.40	0.00	(0.020)	5,216	0.71	1.87	-0.05	(0.051)
# Arrests: Other	1,634	0.17	0.79	-0.01	(0.036)	5,216	1.39	3.01	-0.09	(0.085)
Academics										
In CPS Data	1,634	1.00	0.00	0.00	(0.000)	5,216	0.91	0.29	0.00	(0.008)
Engaged in CPS in June (if ever in CPS)	1,634	0.99	0.12	0.00	(0.006)	4,781	0.51	0.50	0.00	(0.013)
Prior School Year Academics if Enrolled										
Days Attended (if any attendance)	1,629	136.92	30.45	0.70	(1.404)	2,930	122.78	54.28	2.54	(1.823)
Grade (if in school prior year)	1,634	10.15	1.25	-0.04	(0.061)	3,112	10.57	1.12	-0.02	(0.041)
Free Lunch Status (if in school prior year)	1,634	0.92	0.27	0.00	(0.014)	3,112	0.83	0.37	0.00	(0.014)
GPA (if available)	1,574	2.37	0.88	0.00	(0.044)	1,777	1.95	0.96	-0.03	(0.046)
Employment and Earnings										
Has SSN	1,634	0.81	0.39	0.02	(0.019)	5,216	0.71	0.45	0.01	(0.013)
Worked in Prior Year (if has SSN)	1,334	0.07	0.26	-0.02	(0.014)	3,742	0.22	0.41	0.00	(0.014)
Neighborhood Characteristics										
Census Tract: Median Income	1,634	35665	13633	-347	(660)	5,216	33759	13633	-175	(360)
Census Tract: Share HS+	1,634	72.91	15.82	-0.85	(0.79)	5,216	74.00	10.36	-0.36	(0.28)
Census Tract: Unemployment Rate	1,634	19.07	8.66	-0.03	(0.42)	5,216	12.81	4.86	0.14	(0.12)

Notes. 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 tests show treatment coefficients and Huber-White standard errors from a regression of each characteristic on a treatment indicator, randomization block fixed effects, and duplicate application indicators. Test of difference across all baseline characteristics fails to reject the null of no treatment group difference: in 2012, F(63,1545)=1.01 (p=.449); in 2013, F(65,5126)=.78 (p=.897); in the pooled sample, F(69,6709)=.84 (p=.830). Gender not included in table since it is collinear with randomization blocks. 2012 sample was 38.5% male; 2013 sample was all male. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table 2: Summer Program and Non-Program Employment

Program Only	Program + Other Job	Other Job Only	No Job	Program Only	Program + Other Job	Other Job Only	No Job
A. 2012 Cohort							
Treatment (N=603)				Control (N=731)			
71.1%	3.8%	3.8%	16.4%	0.0%	0.0%	15.2%	84.3%
B. 2013 Cohort							
Treatment (N=1913)				Control (N=1829)			
28.9%	3.0%	14.4%	50.9%	0.5%	0.1%	22.6%	74.7%

Notes. Table shows proportion of each group working in the program, outside the program, both, or neither during the summer. Sample includes all youth with non-missing social security numbers (N = 5,076); missing data are balanced across treatment and control groups. Program participation measured with provider records; other work measured by Unemployment Insurance data.

Table 3: Local Average Treatment Effect on Number of Arrests by Year (x 100)

Number of Arrests for:	Total	Violent	Property	Drugs	Other
A. Pooled Sample (N=6,850)					
Year One	-7.48	-6.38***	1.65	2.30	-5.06
	(6.92)	(2.24)	(1.80)	(2.89)	(4.62)
CCM	76.81	18.34	8.2	13.9	36.37
Year Two	0.91	0.78	2.95	-5.25*	2.43
	(6.77)	(1.93)	(1.88)	(2.86)	(4.38)
CCM	57.42	9.52	4.21	18.64	25.04
All Years	-5.76	-5.78*	5.76*	-3.93	-1.82
	(11.45)	(3.33)	(2.99)	(4.56)	(7.38)
CCM	144.0	30.32	12.67	35.47	65.56
B. 2012 Sample (N=1,634)					
Year One	-0.47	-4.18**	1.67	0.61	1.44
	(5.04)	(1.99)	(1.39)	(2.17)	(2.72)
CCM	27.49	9.88	3.11	3.81	10.69
Year Two	2.28	-0.14	3.83**	-2.44	1.02
	(4.72)	(1.69)	(1.74)	(1.87)	(2.64)
CCM	24.01	5.1	1.32	8.14	9.45
Year Three	1.89	0.11	2.83**	-2.73	1.68
	(4.72)	(1.64)	(1.33)	(2.05)	(2.68)
CCM	24.03	5.40	0.67	7.51	10.46
C. 2013 Sample (N=5,216)					
Year One	-13.55	-7.94**	1.72	4.35	-11.68
	(12.04)	(3.75)	(3.09)	(5.01)	(8.21)
CCM	112.13	24.24	11.65	19.97	56.26
Year Two	-1.48	1.33	2.12	-7.82	2.88
	(11.77)	(3.24)	(3.10)	(5.02)	(7.74)
CCM	81.99	12.68	6.41	26.55	36.35

Notes. 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. All regressions estimated using two stage least squares including block fixed effects, duplicate application indicators, and the baseline covariates listed in the text. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table 4: Local Average Treatment Effect 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	0.01	-0.45	0.02	-0.01
	(0.02)	(2.59)	(0.05)	(0.02)
CCM	0.74	91.39	1.95	0.62
N	4993	4993	2447	4993
2012	0.00	-3.16	-0.05	-0.01
	(0.01)	(2.63)	(0.05)	(0.02)
CCM	0.95	133.39	2.27	0.88
N	1427	1427	1218	1427
2013	0.01	1.43	0.13	-0.02
	(0.04)	(4.39)	(0.13)	(0.04)
CCM	0.58	57.98	1.37	0.42
N	3566	3566	1229	3566

Notes. Table 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 (other missing data treatments shown in appendix). 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 estimated using two stage least squares including block fixed effects, duplicate application indicators, and the baseline covariates listed in the text. CCM indicates control complier mean. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table 5: Local Average Treatment Effect on Formal Employment Outcomes

Outcome:	Any Formal Employment	Any Provider Employment	Any Non-Provider Employment	All Earnings
		Panel A. Pooled Sample (N=5,076)		
During Program	0.85***	1.04***	-0.06**	1013.54***
	(0.03)	(0.01)	(0.03)	(81.08)
CCM	0.12	0.00	0.16	122.66
Remaining Year One Quarters	0.03	0.04***	-0.01	67.57
	(0.03)	(0.01)	(0.03)	(167.44)
CCM	0.22	0.00	0.22	582.76
Year Two	0.02	0.09***	-0.02	155.82
	(0.03)	(0.01)	(0.03)	(252.97)
CCM	0.44	0.04	0.40	1237.47
		Panel B. 2012 Sample (N=1,334)		
During Program	0.88***	1.07***	-0.07***	1246.41***
	(0.03)	(0.01)	(0.02)	(91.61)
CCM	0.10	0.00	0.16	333.69
Remaining Year One Quarters	-0.06**	-	-0.06**	-213
	(0.03)		(0.03)	(146.65)
CCM	0.22		0.22	672.81
Year Two	-0.01	0.04**	-0.03	-173.14
	(0.03)	(0.02)	(0.03)	(180.56)
CCM	0.42	0.04	0.38	1215.47
		Panel C. 2013 Sample (N=3,742)		
During Program	0.82***	1.02***	-0.04	781.69***
	(0.04)	(0.02)	(0.04)	(128.34)
CCM	0.15	0.00	0.16	19.36
Remaining Year One Quarters	0.11**	0.07***	0.04	324.72
	(0.05)	(0.01)	(0.05)	(290.57)
CCM	0.21	0.00	0.21	468.97
Year Two	0.06	0.14***	-0.01	487.18
	(0.05)	(0.02)	(0.05)	(455.60)
CCM	0.45	0.04	0.42	1170.13

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 earnings are imputed with 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. All regressions estimated using two stage least squares including block fixed effects, duplicate application indicators, and the baseline covariates listed in the text. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table 6: Multiple Hypothesis Testing Adjustments

	Program Effect		Adjusted P-Value			
	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. 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
D. 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
E. 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

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 10,000 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 7: Heterogeneity in Local Average Treatment Effect by Predicted Treatment Impact

	Any Post-Program Formal Employment	Number of Violent Crime Arrests	School Persistence
Largest Quartile of Predicted Responders	0.14**	-10.34	0.04
	(0.06)	(9.05)	(0.05)
Rest of Sample	-0.01	-4.24	-0.02
	(0.04)	(3.34)	(0.02)
P-value, test of subgroup difference	0.02	0.53	0.22
CCM Largest Quartile of Predicted Responders	0.34	54.97	0.60
CCM Rest of Sample	0.53	21.00	0.62
N	5076	6850	4993

Notes. Any post-program employment equals 1 if the youth had any UI-covered employment in the 6 post-program quarters, defined for anyone with non-missing employment data. Number of violent crime arrests calculated over the entire follow-up period (2 years for 2013 cohort and 3 years for 2012 cohort) for all observations. School persistence defined for those who ever appear in the CPS data and had not graduated prior to the program. Persistence equals 1 if the youth either graduated after 2 post-program school years or continued to attend during the third post-program school year. Table shows the LATEs for the dependent variable at the top of each column by subgroup. Subgroups defined by the causal forest predictions for each outcome. The largest quartile of predictions is the most positive for employment and school persistence and the most negative for violent-crime arrests. Standard errors clustered on individual. All regressions estimated using two stage least squares including block fixed effects, duplicate application indicators, and the baseline covariates listed in the text. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Table 8: Summary Statistics by Quartile of Predicted Employment Impact

	Quartile of Predicted Employment Impact			
Variable	Q1	Q2	Q3	Q4
Prediction, Any Post-Program Employment	-0.033	-0.004	0.019	0.052
Take-Up Rate	0.403	0.409	0.424	0.453
In 2012 Cohort	0.203	0.225	0.244	0.380
Age at Program Start	18.628	18.474	18.038	16.889
Hispanic	0.008	0.020	0.065	0.162
Male	0.883	0.856	0.849	0.765
Any Baseline Arrest	0.585	0.546	0.512	0.306
Graduated Pre-Program	0.340	0.303	0.240	0.100
Engaged in CPS in June	0.455	0.481	0.623	0.849
Days Attended in Prior School Year (if any)	110.253	118.473	122.386	138.985
GPA	2.007	2.146	2.081	2.196
Worked in Prior Year	0.179	0.229	0.225	0.101
Census Tract Unemployment Rate	17.364	14.702	12.841	12.291

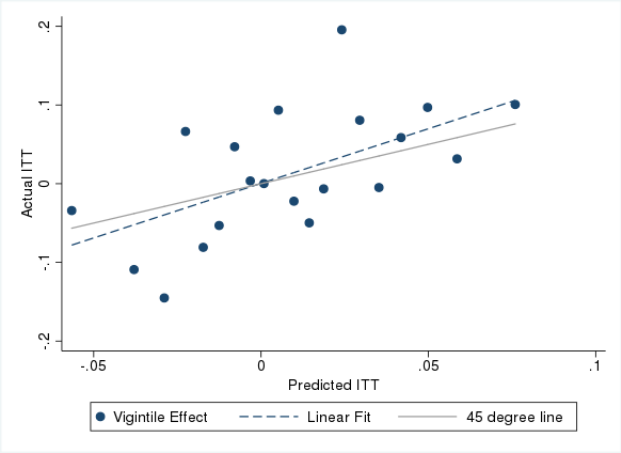
Notes. Table shows mean baseline characteristics for each quartile of predicted employment treatment impacts. Predictions from a causal forest.

Table 9: Heterogeneity in Local Average Treatment Effects across Outcomes by Predicted Employment Response

	Any Provider Employment	Any Non-Provider Employment	Post-Program Earnings	School Persistence	Social Cost of Crime
A. Employment, School Persistence, and Social Costs of Crime					
Largest Quartile of Predicted Employment Responders	0.11***	0.1	446.22	0.08*	-13303.18
	(0.03)	(0.06)	(497.03)	(0.04)	(12558.31)
Rest of Sample	0.11***	-0.06*	134.05	-0.04	-8258.91
	(0.02)	(0.04)	(484.82)	(0.03)	(7177.44)
P-value, test of subgroup difference	0.93	0.02	0.65	0.02	0.72
CCM Largest Quartile of Predicted Employment Responders	0.04	0.30	988.25	0.73	45284.9
CCM Rest of Sample	0.04	0.50	2117.38	0.60	50247.71
	Total	Violent	Property	Drugs	Other
B. Number of Arrests by Type					
Largest Quartile of Predicted Employment Responders	8.53	-5.65	10.05**	0.07	4.06
	(19.15)	(6.24)	(4.52)	(7.75)	(11.96)
Rest of Sample	-8.84	-7.88*	2.97	-2.75	-1.18
	(15.08)	(4.49)	(3.87)	(6.25)	(9.71)
P-value, test of subgroup difference	0.46	0.77	0.23	0.77	0.73
CCM Largest Quartile of Predicted Employment Responders	105.28	26.17	5.25	25.67	48.18
CCM Rest of Sample	157.72	34.06	17.47	37.48	68.72

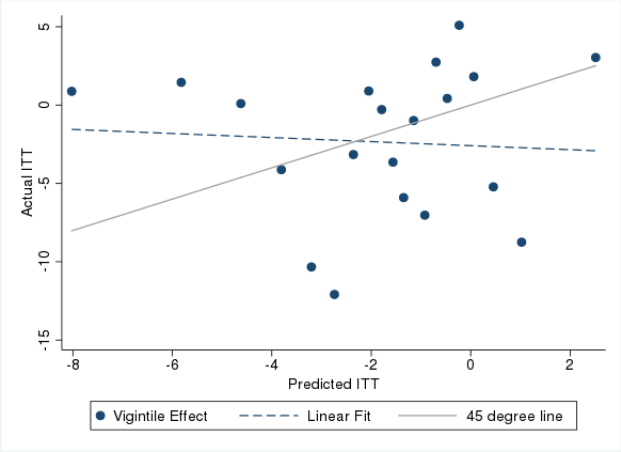
Notes. Table shows the LATEs for the dependent variable at the top of each column by predicted employment response (largest quartile of predicted employment response versus quartiles 1-3). Sample restricted to youth with non-missing employment data (n=5,076). Employment outcomes defined over the 6 post-program quarters. School persistence equals 1 if the youth either graduated after 2 post-program school years or continued to attend during the third post-program school year. Persistence sample additionally limited to youth who had not graduated prior to the program (N=3,829). Arrest counts and associated social cost use 3 years of post-randomization arrests for 2012 cohort and 2 years for 2013 cohort. All regressions estimated using two stage least squares including block fixed effects, duplicate application indicators, and the baseline covariates listed in the text. Standard errors clustered on individual. Stars indicate: * p<0.1, ** p<0.05, *** <0.01.

Figure 1: Predicted versus Actual Effects, Post-Program Employment



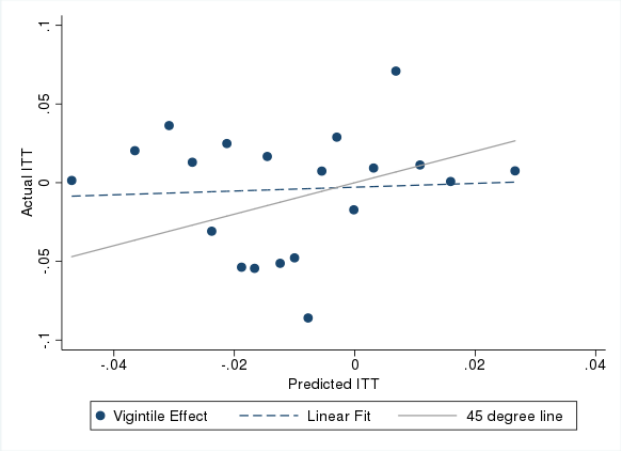
Notes. Figure shows vigintiles of predicted intent-to-treat effects on post-program employment versus the actual estimated effect for individuals predicted to be in each vigintile. The dashed line shows the linear relationship between the actual and predicted effects. The solid line is the 45-degree line which is included for reference.

Figure 2: Predicted versus Actual Effects, Cumulative Violent-Crime Arrests



Notes. Figure shows vigintiles of predicted intent-to-treat effects on cumulative violent-crime arrests versus the actual estimated effect for individuals predicted to be in each vigintile. The dashed line shows the linear relationship between the actual and predicted effects. The solid line is the 45-degree line which is included for reference.

Figure 3: Predicted versus Actual Effects, Persistence in School



Notes. Figure shows vigintiles of predicted intent-to-treat effects on persistence in school versus the actual estimated effect for individuals predicted to be in each vigintile. The dashed line shows the linear relationship between the actual and predicted effects. The solid line is the 45-degree line which is included for reference.