Appendix Material for When Scale and Replication Work: Learning from Summer Youth Employment Experiments

Sara B. Heller (University of Michigan & NBER)

April 14, 2021

A Study Design

A.1 One Summer Chicago Plus 2015

OSC+ has been studied in two previous cohorts, 2012 and 2013. The program itself has changed somewhat over time. In the initial 2012 cohort, 14- to 21-year-old high school students spent a total of 8 weeks working in a mix of non-profit and government jobs after a 1-day training. They worked up to 15 hours a week and participated in a cognitive behavioral therapy-based social-emotional learning (SEL) curriculum for another 10 hours per week. Everyone was assigned an adult mentor to provide support and help them deal with barriers to employment throughout the summer. The second year of the program (2013), the program was only open to boys 16- to 22-years old. Half were recruited from high-violence high schools, and half were recruited through agencies connected to the criminal justice system. The program was shortened to 6 weeks, and private sector jobs became part of the program. The shifts in the 2015 program are described in the main text: the population expanded; recruiting was again school-based; girls were included again; the SEL curriculum changed to a civics curriculum and was shortened; and the wages dropped relative to the labor market.

In spring 2015, OSC+ providers advertised the program and solicited applications from youth at 49 high schools in high-violence community areas. Initially, 5,444 applications were collected. We grouped applicants into geographic blocks based on the address listed on their application, because one of the main concerns about prior lotteries had been that some youth were required to make long commutes to work. We then randomly ordered applicants within block and assigned them to both a provider serving that area and one of three treatment groups: job only, job + mentor (including a Civic Leadership Foundation curriculum delivered on Fridays), and a control group.

Through the process of matching these applications to administrative data, we identified 39 cases where the same youth submitted more than one application, with identifying information different enough that we did not catch the second application in our initial de-duplication process. We collapse duplicate applications into a single observation, where treatment is defined as the maximum of all random assignment indicators (since the max of two random variables is also random). To ensure the analysis accounts for the higher probability of treatment among these applicants, we include dummy variables for having submitted two applications in all analyses.

A.2 WorkReady Experimental Design 2017

During the study years, WorkReady[35] had 3 program models:

1. *Service learning*: Youth worked in groups to address a community problem. They researched the problem, developed solutions, and provided direct service and advocacy to address the problem. This model was designed for youth with minimal or no prior work experience.

2. *Work experience*: Youth participated in a structured work experience, with an emphasis on skill development and ongoing adult interaction. Youth were placed in a professional work environment appropriate for those with little or no work experience. They received training in workplace skills prior to the start of their job and completed work-based learning projects.

3. *Internship*: Youth who were already prepared for the workplace were placed in summer jobs, with a focus on placements that would not otherwise be accessible to young people. They also participated in professional development workshops and interacted with a trained adult supervisor over the course of the summer.

[35] Note that there is a research report on a Philadelphia summer jobs program implemented in the late 1990s, but that was a different program was run by a different organization (McClanahan et al., 2004). Although that study was intended to be a randomized controlled trial, the appendix reports dropping all control youth who actually participated, then estimating the difference between participants and non-participants (not treatment and controls). So to our knowledge, there are no experimental estimates of any Philadelphia summer jobs programs.

Typically, providers assigned youth to a program model based on their age and experience. Youth interested in WorkReady applied in one of two ways: through a program provider or through PYN's online system without any attachment to a particular provider. Youth applying through providers have often already been screened to meet a given provider's program requirements, which vary across providers. Youth applying without a connection to a particular provider entered a group called the "general pool." In the absence of a lottery, PYN tries to place these general pool youth in positions that open up as providers request more youth to fill their slots. Because this group is often more disadvantaged and less likely to get a program slot, PYN wanted to prioritize increasing access among this group. At the same time, many providers have strong preferences about whom they are willing to serve (sometimes driven by requirements of the jobs, e.g., passing criminal background checks), and so do not accept youth they have not screened themselves. To achieve the dual goals of expanding access to the program to youth who might not otherwise be served and allowing providers to screen youth according to their needs, we allowed providers to choose one of the two types of lotteries described below. A handful of providers were exempted from the lottery by preference or for logistical reasons, but 59 providers participated. We conducted two waves of randomization. The first occurred on May 1, 2017, and the second occurred on May 12, 2017.

Over-Recruitment In the over-recruitment lotteries, individual program providers collected applications from more youth than they could serve. They were able to pre-screen applicants to meet their own operational requirements, but were asked to collect more applications of acceptable candidates than they had program slots. Part of the rationale from their perspective was that the City of Philadelphia was trying to scale up the program, so reaching out to new youth would inform these youth about the program and establish relationships for the future, even if they did not win a program slot in this program year.

Providers were given targets for the number of applications to submit relative to the number of available slots, but the exact number of applicants relative to those targets varied somewhat by provider. They were able to choose which youth to place directly in their non-lotteried slots, and which to submit to be randomized for the remaining open slots. Providers submitted their list of applicants to PYN, and the researchers randomized within provider.

Since there are a number of steps, including paperwork completion, that generate drop-off between the program offer and actually participating, we aimed to randomly select about 120 percent times the number of available slots within each provider as treatment youth. The remainder were assigned to the control group. As a result, treatment probabilities varied across providers. Overall, 1,554 applicants took part in an over-recruitment lottery across 39 providers, with 765 assigned to treatment.

General Pool Some program providers did not need to prescreen applicants and were willing to accept youth unknown to them who applied through the general pool. To ensure that program providers could logistically serve the general pool youth, providers submitted general location and age preferences (region of the city by zip code and whether they could serve 14- to 16-year olds, 17- to 18-year olds, or both). We individually randomly assigned general pool applicants within these age and location blocks to providers based on how many program slots those providers reported needing to fill. As in the over-recruitment lottery, we randomized about 120 percent more applicants to treatment than requested slots to ensure that all slots could be filled, even when youth were hard to find, failed to complete paperwork, or were no longer interested. Within age and geography blocks, provider assignment was random. Although the providers who were in these blocks are not necessarily representative of all providers, this does induce some random variation in what type of program youth were assigned. Overall, 1,838 general pool applicants were randomly assigned across 20 providers, 571 to treatment and 1267 to control.

A.3 WorkReady Experimental Design 2018

To simplify the experimental design in 2018, there was only one lottery consisting of youth who applied to the general pool. PYN learned which providers were willing to accept general pool youth and assessed how many slots these providers would need to fill after having submitted all their paperwork on the youth they recruited directly. PYN also worked to collect all the necessary paperwork from the general pool youth, so that the youth would be ready to place (therefore reducing barriers to providers serving treatment youth).

As of the beginning of June, there were 1,105 general pool applications that had not been placed in other jobs, and providers anticipated needing an additional 344 youth to fill their program slots. Because we knew from prior years that not all treatment youth would participate, we randomly assigned 450 youth from the 1,105 applicants to be in the treatment group. PYN then worked to place these treatment youth at about 40 different providers according to provider needs. As a result, in this cohort, there was no random variation in provider.

B Data

All data come from probabilistic matches to government agency data, accessed by legal agreements with the agencies. Researchers performed the match to Chicago Police Department and Chicago Public Schools data, and staff at each agency performend the match at the Philadelphia Police Department and School District of Philadelphia.

For the researcher-performed match in Chicago, we use information on name, date of birth, gender, race, and home address across program records, CPD, and CPS records to identify which rows of each database belong to the same person. Each individual database includes an individual identifier that links records on the same person, but the identifier can include some error (e.g., if a student transferred schools and was assigned a new student ID number rather than correctly transferring the existing number). The probabilistic algorithm allows for typographical or categorical errors (e.g., mismatches on race or gender) both within data sources and when linking across databases.

To minimize false positive links, we impose logical constraints during the record linkage process. If a potential link between two records would violate a prespecified constraint, this link is not permitted. Specifically, we do not link records if the resulting individual would contain school records such that the total number of attendance days in a given year exceeds 200 days, or such that a student would have appeared to take the same standardized test twice within the same year.

Because treatment may affect how many records an individual has during the post-randomizaton period, it is theoretically possible that treatment could be correlated with how likely observations are to be linked. For example, if the program changes the probability of arrests, we are more likely to observe arrest data for individuals in one group compared to the other, which means we potentially have more identifying information about individuals in that group. If we run the record linkage algorithm with pre- and post-randomization data combined, it is possible that a pre-randomization arrest record may only be linked to a person through a post-randomization arrest record.

To avoid injecting treatment effects into linkage probabilities, we run an initial match using only pre-randomization records. After the pre-randomization record linkage is complete, we then re-run the record linkage process with post-randomization school and arrest records included. Records linked in the first step cannot be broken up; that is, post-randomization information can not change the decision about whether pre-randomization records belong to a given study individual. But new records can be added to the existing matched individual, and non-matched individuals can be newly matched to their post-randomization data. This approach helps ensure that treatment does not influence the probability that given records are correctly linked to the right person.

B.1 Chicago

Chicago data were used via data agreements with the University of Chicago Urban Labs, where all data are stored and analyzed. Variables were defined in a similar way to prior work on OSC+, so the data appendix in Davis and Heller (2020) largely applies here as well. The primary data sources for Chicago are school records from Chicago Public Schools (CPS), arrest records from Chicago Police Department (CPD), and the OSC+ study participant file. Records in all files include information about name, date of birth, gender, race, and home address. In the Chicago Public Schools data, we observe historical enrollment records, including reasons for leaving, that capture everyone who ever enrolled in the public school district since the early 2000s. We have data through the 2016-17 academic year, or 2 post-program school years. These enrollment records can be easily deterministically linked via a student identifier to other schooling data, such as grades, attendance, and standardized test scores. Of the initial 5,444 application records (prior to deduplication, see above), 99.5 percent matched to CPS data. The remaining 26 applications may not have matched because they were enrolled in private or parochial schools or schools outside of Chicago, or because the fields used to match were not sufficiently similar between the OSC+ and Chicago Public School data. As described in the main text, we exclude these youth from the main education results, since we have no information on their school outcomes. They receive different imputations below in the section that assesses how missing data affects schooling results.

In the Chicago Police Department data, we observe historical arrest records, including the statute under which the arrest was made, that capture everyone who has been arrested by Chicago Police Department since 1999. Anyone not matched to the arrest data receives a 0 for all arrest counts. Crime categories are the same as in our prior work on OSC+.

B.2 Philadelphia

B.2.1 CARES Data

The City of Philadelphia's Data Management Office maintains an integrated data system (called CARES) that collects and integrates service records across multiple city service agencies. They received our personal identifiers from PYN application information, including name, date of birth, and gender. The Data Management Office then used probablistic matching to find each youth in City service data; the Philadelphia Police Department also provided them with full arrest databases to which they could match our data. In practice, the threshold the City used as a cutoff for probablistic matching was high enough that when using only name, date of birth, and gender as matching variables, the match had to be exact to exceed the threshold. However, the process did still allow for typographical errors in records in two ways. First, they match their own records across agencies using more identifiers than we had available in our data. If two records had similar addresses and parent information but different spelling variants or DOB typos, for example, they could still match to each other. Our WorkReady record would then only have to match to one of the records to be considered a match; the record with other variants would still be linked to our study individual.

Second, as described below, the City sent identifiers and their own study indentifying numbers to the School District, so that we could merge education data back into City data despite the records being stripped of identifiers. The School District used a different probabilistic algorithm that caught some additional matches. Before the City pulled the final data, they used the School District match to reconcile the conflicting match cases. So in practice, our identifiers could match to City data despite not being exact matches on name, date of birth, and gender. As with any matching, it is possible that the matching process for these outcome measures missed some name or date of birth variants and therefore understate true service receipt (since we assign 0s for those not matched to the data). Importantly, we used personal identifiers that they youth provided on the application, prior to random assignment. This means that there is no reason to think that typographical errors or nicknames preventing exact matching would be imbalanced across treatment and control groups.

The City data team stripped the identifying information off of the data and returned it to us identified only by a study ID number. We were required to securely delete all our application records prior to receiving the data, so that there was no way to re-identify the records. The City is keeping our initial match file as a record of study participants. They also provided a separate file with HIPAA-covered records (mental health and substance abuse services). For these matches, we provided a file with identifiers attached to a smaller number of covariates, so that no individual was uniquely identified by any combination of covariates. These data have different study identifying numbers and so can not be merged with other records.

B.2.2 School District Data

The School District of Philadelphia (SDP) received identifiers from the City and used their own probabilistic matching algorithm to link the records to school data. SDP has enrollment and graduation records for every public school student, but charter schools do not typically report grades, absences, or suspensions to the District office. Additionally, their enrollment file only records enrollment status in the current academic school year (unlike Chicago where enrollment records are cumulative). Because our data agreement for this project only covered data starting in the 2016-17 academic year, the earliest we can observe whether a study youth is enrolled, dropped out, transferred, or graduated, is that academic year (the year prior to randomization for the 2017 cohort and 2 years prior for the 2018 cohort).

We use the pre-randomization records to identify youth who graduated prior to the program (n = 286). Since they can not have changes in education outcomes by construction, we exclude them from all education analysis. Another 353 youth did not match to SDP data at all and so are missing all school data. These youth may not have matched because they were enrolled in private or parochial schools or schools outside of Philadelphia. It is also possible that some of the youth were enrolled in SDP, but with identifiers different enough from what they used on the WorkReady application that the matching process missed them. We exclude these youth from the main school persistence outcome, since we do not have information on whether they are enrolled or graduated at any given time, but they are included in the imputation robustness checks in Appendix Section C.4.

Another set of youth do appear in the SDP data, but are missing individual variables or years of data for multiple reasons. Charter school students do not typically have grades, absences, or suspensions, although some charters report those variables in an irregular and unreliable way. We therefore set those variables to missing for charter students.[36] Students who drop out, transfer, or graduate will be missing data for some years, as will those who become incarcerated, deceased, or otherwise incapacitated. The fact that there is so much missing data, and the fact that the reason for missingness varies – sometimes indicating school success, as in the case of graduation, sometimes indicating school failure, as in the case of dropout, and sometimes indicating a neutral reason like transfer – makes it difficult to draw strong conclusions from educational outcomes.

In the main text, we focus on the outcome we can measure for everyone other than the 353 non-matches and 286 prior graduates: whether a student is still in school or has graduated by a given year. Section C.4 below presents results under different imputation schemes for the missing data, including the extreme assumptions that all missing data is an indication of either school success or school failure.

[36] There is one case where a student is enrolled in a District school prior to randomization but has absence data from a charter school. This results in the student's absence record being missing despite having non-missing grade and suspension information. The student graduated prior to the program, so this does not change any results. But it does explain why the baseline N for suspensions and attendance do not match exactly.

C Robustness Checks

C.1 Covariates

In the main analysis, baseline covariates are included in all the regressions. For non-HIPPA data in Philadelphia, covariates include dummy variables indicating age bins, gender, race/ethnicity, any receipt of baseline services by type, number of baseline arrests by type, and prior grade level. For HIPAA data, the baseline covariates are gender, an indicator for being Black, age bins, and indicators for prior receipt of HIPAA-covered services. In Chicago, we include indicator variables for male, race, age bins, number of baseline arrests by type, GPA categories, number of absence bins, and duplicate randomization. To reduce any potential finite-sample misspecification, we have included baseline covariates as dummy variables. Missing baseline covariates are imputed as 0, with missing data indicators included.

Tables A1 and A2 show all the main results without baseline covariates, other than the block and duplicate indicators needed to ensure treatment is conditionally random. None of the substantive conclusions change relative to the regressions including baseline covariates in the main text. Because there is a slight imbalance favoring the treatment group on a couple baseline variables in Philadelphia, results excluding covariates tend to be larger and more statistically significant there. In Chicago, some results that are right on the border of standard significance cutoffs cross the cutoff without covariates, most becoming more significant. But overall, the findings of both studies are robust to excluding baseline covariates.

C.2 Alternative Functional Form

The dependent variables in the paper are generally either indicators or counts. Tables A3 reports the results of either probit or Poisson regression (with robust standard errors to relax the assumption that mean equals variance) to ensure findings are not sensitive to the functional form of the regression. We note that for WorkReady, no observations had more than 1 drug or other arrest during year 1, so those results are from a probit rather than Poisson regression. All coefficients are reported as average marginal effects.

To ensure convergence, the non-linear regressions are run with a limited set of baseline covariates. For Philadelphia non-HIPAA outcomes and Chicago, this includes indicators for Black, male, and whether someone had a baseline arrest. For Philadelphia HIPAA outcomes, it includes indicators for Black, male, and whether someone had any behavioral health service prior to baseline.

Given the rarity of most outcomes in Philadelphia, many randomization blocks are dropped due to lack of variation in the dependent variable within strata. As such, the tables report the N used in the regression, as well as the control mean specific to the sample contributing to identification. The overall pattern of results shows extremely similar changes to the main results. In Philadelphia, the proportional changes are slightly bigger than the linear probability models in the main text. This may be because probit handles the low base rates more effectively. It may also be because the regressions include fewer baseline covariates to ensure convergence, and covariates help to control for the small amount of baseline imbalance. In Chicago, the main difference is that results are slightly less precise due to the inclusion of fewer baseline covariates.

C.3 Randomization Inference and Multiple Testing Adjustments

At the time of random assignment for WorkReady, we set a balance check rule of re-randomizing until each individual covariate was balanced at the $p>0.1$ level and the joint test of balance had $p>0.05$. None of the randomizations required more than one draw of random assignment vector to meet this rule. But since this is the rule we would have followed, our intention was for our randomization inference to throw out any potential random assignments that would not have met this rule (Morgan and Rubin, 2012). However, we were legally obligated to delete all our application data prior to receiving de-identified outcome data, and not all of our previously available data was transferred back to us for the outcome analysis. As a result, we no longer have all the information on which we tested balance initially. Since we can not implement the exact randomization rule, we use straightforward randomization inference that does not exclude any randomization vectors. This should be conservative, making our adjusted p-values slightly larger than they would otherwise be.

We implement standard randomization inference separately for each main outcome, which tests the sharp null of no treatment effects for anyone. This test is more robust to outliers and clustering than standard tests of the null of no average treatment effects, and it more directly ties the statistical test to the randomness induced by random assignment. We re-randomize 5,000 times, reporting the probability we would find the treatment coefficient as large (in absolute value) as the one in the actual data given the distribution of treatment coefficients under the null.

Since the probability of a Type I error increases with the number of hypothesis tests performed, we also adjust our inference to control for the number of tests run within families (the WorkReady PAP pre-specifies that we would adjust within families). We control both for the family-wise error rate (FWER), which ensures that the probability we reject any null within the family of tests remains less than α, as well as the false discovery rate (FDR), which relaxes the conditions for rejection in exchange for additional power by allowing for q percent of null rejections to be false (Westfall and Young, 1993; Benjamini and Hochberg, 1995). For the FWER adjustment, we use the Stata command wyoung, written by Julian Reif (which uses bootstrap-based resampling rather than permutation of the treatment indicator; we use 5,000 bootstraps). For the FDR adjustment, we use Michael Anderson's code based on the original Benjamini and Hochberg approach. We report the adjusted p-values under FWER control and the q-value – the smallest proportion of false rejections we could allow and still reject each null. See Davis and Heller (2020) and Heller et al. (2017) for additional discussion of the pros and cons of each adjustment.

To perform these adjustments, we group our hypotheses into families to answer the following questions: 1) do any criminal justice outcomes move? (any juvenile incarceration, any juvenile justice services, total number of arrests), 2) what type of crime changes? (violent, property, drug, and other arrests), 3) do family outcomes move? (fertility, receipt of child protective services, homeless shelter use), and 4) do behavioral health outcomes move? (receipt of substance abuse services, receipt of mental health services). Conceptually, we might prefer the fertility outcome to be grouped with behavioral health to capture risky behavior. But because our legal agreements required all HIPAA-covered services to remain in a separate file with limited covariates to avoid re-identification (despite the fact that the data do not include personal identifiers), we can not combine those outcomes into a single family. We make no adjustment to school persistence, since it is the single outcome in the education family.[37] Note that family 2 is the only relevant family for the OSC+ study given the data available for Chicago.

Tables A4 and A5 show the p-values from the original robust standard error calculation, p-values from randomization inference, and adjusted p- and q-values under FWER and FDR control for WorkReady and OSC+ respectively. Randomization inference changes little; p-values move around a small amount, but substantive conclusions are basically unchanged. Adjusting for the number of tests, on the other hand, often pushes the probability of false rejections above traditional cut-offs. The decline in other arrests in Philadelphia is the only result that remains at or below $p = 0.05$ across all the adjustments.

If this were the first and only set of tests for SYEP effects, these results might merit considerable caution about drawing conclusions. But there are several arguments for resisting too much caution here. First, the fact that we have two entirely independent samples across the two studies provides a built-in replication. The fact that both drug and other arrests decrease in both settings is informative; the probability we would be falsely rejecting both nulls across two independent experiments is much lower than the probability that any single result is a false rejection. Second, the decline in arrests and incarceration has also occurred in every other study of SYEPs where they have been measured. This suggests that a null hypotheses of no effects might actually be conservative; all of the existing evidence suggests that these effects are unlikely to be zero. We are therefore inclined to think that, at least for criminal justice outcomes, the adjustments for multiple testing often move p-values above the typical 0.1 threshold largely because the non-compliance in the study reduces power, not because the probability of false rejections is actually high.

We should be more cautious about the result that child protective services decline. That outcome is only available in Philadelphia, so we lack the built-in replication across cities. And it is not an outcome that has been tested elsewhere, so zero-impact is a more reasonable null hypothesis. Because it is measuring an extremely costly outcome — a substantiated risk of child abuse and neglect — the possibility that SYEPs affect the outcome is worth some attention even if there is some risk of a false rejection. And the fact that there is a more

[37]This is a departure from our pre-analysis plan for Philadelphia, driven by the unexpected amount of missing data from charter schools. It is, however, consistent with our prior analysis of OSC+ in Davis and Heller (2020). We show other outcomes without multiple testing adjustments in Appendix Section C.4.

precisely-estimated decline among Black youth, who have elevated rates of these services in the control group, is worth noting, although certainly not dispositive given how many tests are run in the subgroup analysis. As we argue in the main text, the result should be a priority for future research before we draw strong conclusions about SYEPs effects.

C.4 Imputed School Data

As discussed above, there are multiple reasons for missing school data. Tables A6 through A9 show education outcomes with different assumptions about missing data, separately for each study's ITT and LATE. All regressions drop pre-program graduates, since they can not have school data by construction. We make one departure from our pre-analysis plan in these outcomes by excluding Keystone test scores as an outcome. This is because the tests were not regularly administered to everyone in the sample in every year, which we did not know at the time of pre-specification.

In each table, Panel A shows the treatment-control difference for absences, GPA, and suspensions for non-missing data only, which implicitly assumes data are missing completely at random. For students with at least 1 day attended in a school year, we assign 0s for suspensions and absences if those variables are missing. Panel B imputes the mean outcome variable by group (treatment and control) and study year for all missing data. This assumes that the data are missing completely at random conditional on study year and group. Panels C and D make the extreme assumption that all missing data is missing for either the most low-performing or most high-performing students, respectively, in the spirit of Lee bounds. For the low-performing imputation, we assign a 0 GPA, 1 for ever being suspended, and the 95th percentile of non-missing days absent for all missing observations. For the high-performing imputation, we assign the 95th percentile of GPA, 0 for ever being suspended, and 0 days absent for all missing data. The final two panels take a hybrid approach. They assume that data are missing completely at random (conditional on year and group) for those who are marked as transfering out of the district in the school records, and assign transfers the group-year mean. They then assign all-low or all-high values for the remaining missingness.

For WorkReady, the fact that the control means move around so much across imputations emphasizes how much missing data there is. This is largely from the fact that about a third of youth attend charter schools, which do not report grades, misconduct, or days absent to SDP. Because the treatment-control difference on missing data is relatively small, the various imputations do not change the substantive conclusion that there is little significant change in school outcomes. If anything, there is perhaps an indication of small decreases GPA and days absent, and small increases in suspensions, though no results are statistically significant. But the amount of missing data makes it hard to draw clear conclusions, since the point estimates do move around under the different assumptions about missing data.

OSC+ has less missing data, in part because charter schools in Chicago do report attendance data, though not grades. The coefficient on GPA moves around somewhat less as a result, always indicating negative but quite small and statistically insignificant effects on GPA. The coefficient on days absent moves around more across the various imputations, flipping sign but never differentiable from zero. Ever being suspended, on the other hand, consistently increases by between 8 and 11 percentage points for compliers, which is proportionally quite a large increase relative to complier means (between 51 and 73 percent). It is not entirely clear how much stock to put in this result; prior work on OSC+ has not reported misconduct results, because the school district has warned how unreliable the data are (with reporting quite inconsistent, even within schools). And no other SYEP study has found an increase in misconduct.

It is certainly possible that this is a real treatment effect; as shown in Table A12 below, there is also an indication that treatment youth are less likely to still be enrolled or graduate in the second post-randomization year. Table A10 shows that these negative impacts may be concentrated in the jobs-only treatment arm, although the arms are not statistically different from each other. One possibility is that, by reducing arrests, the program keeps some youth who would otherwise be incarcerated in traditional schools, rather than the schools in juvenile detention or prison. If so, they might be less likely to continue attending (whereas attendance in criminal justice facilities is mandatory), and more likely to commit disciplinary infractions. Or, given the slight concentration of these effects in the job-only arm, it is possible that the stronger connection to the labor force introduces additional problems at school. This is certainly an issue to which future work should attend. But given the uncertainty about data quality and the fact that the result is uncommon across SYEP studies, we are unsure how confident to be that this is a real treatment effect.

D Longer-Run Results

Table A12 shows year 2 results for the 2017 cohort of WorkReady (the only one observed long enough in the data to have year 2 results), and results for years 2 and 3 for OSC+. Since the year 2 results are only a single cohort, they should be compared to just the 2017 cohort's results in year 1, shown separately in the next section.

In both cities, there is some suggestion that crime reductions last beyond the first post-randomization year. In Philadelphia, property crime declines by 2.2 arrests per 100 compliers, effectively eliminating property crime relative to the control complier mean. Point estimates on total and other arrests remain proportionally large and negative, but not statistically significant. Similarly, point estimates on mental health and child protective services are negative and proportionally large, but not significantly different from 0. In Chicago, year 2 arrest results are typically negative but not significant. But in year 3, the drops in total, property, and other arrests are all marginally significant again.

The education results are less encouraging. There is a marginally significant decline in school persistence in OSC+ of 6.2 percentage points. As discussed in the previous section, the results by treatment arm suggest this is driven by the jobs-only group. So it is possible that a summer job unaccompanied by additional program supports gives youth more of a chance to engage with their employer, pulling them into the labor market and out of school. Explanations not specific to treatment arm are also possible. It could be that summer jobs are pulling youth out of summer school, making it harder for them to complete the credits they need to motivate staying in school. Or it could be that the decline in crime keeps youth out of detention facilities, which force them to attend school (but rarely give them enough credits to graduate). If youth who would otherwise drop out do not get incarcerated, they are more free to unenroll in school. This could decrease enrollment at the margin, even if it does not change eventual educational attainment.

If this decline in school engagement persists, it will be an important caveat to the overall effects of SYEP. But it is also in conflict with findings in other contexts – Philadelphia here, where there is no change in school persistence, as well as zero school effects in prior Chicago and NYC work, and positive impacts in Boston. Given that many of the youth are not yet old enough to complete their school career, it will be important to follow up after more time has passed to see if these results persist.

E Subgroup Interactions

For WorkReady, the subgroup divisions included here (gender, race, and age) were pre-specified in the pre-analysis plan. We added one additional split – cohort – ex post, because it adds useful, if unanticipated, variation in baseline risk (the 2018 cohort turned out to be at far lower risk of negative outcomes, helping to identify variation in risk that is useful for our heterogeneity analysis). See Table A11 for baseline descriptive statistics by cohort. We also exclude one split, which the pre-analysis plan specified was more of interest for program implementers than for substantive reasons (whether the youth had previously participated in certain types of City programming). For OSC+, we did not post a pre-analysis plan, since the analysis was intended to follow the previous analyses of OSC+ quite closely. As a result, we focus on the subgroup splits that were previously reported in Davis and Heller (2020): gender and prior arrests. The previous paper also reported splits on whether youth were in school and whether they had worked before. But this 2015 study population was all in school prior to the program, so that split does not apply here. And we do not have access to data on employment records, so we can not test the prior employment split.

We hesitate to over-interpret subgroup differences, because they involve a huge number of hypothesis tests, and because we often lack the statistical power to differentiate between groups. These splits serve mostly to generate variation in baseline rates of the outcomes, used in the main risk-responsiveness analysis. Nonetheless, this section highlights some of the results, which should be considered tentative until validated in future studies that are better powered for this kind of analysis. All estimates come from a model with a one-way treatment interaction with the relevant covariate, but the linear combination of coefficients and standard errors, representing the net effect for each group, are shown to ease interpretation.

Table A13 shows that the 2018 cohort of WorkReady consistently faces a floor effect across many outcomes. The 2018 control group hits the floor on several outcomes, with 0 other types of arrest, no parenthood, and no homeless shelter use. Because the 2018 control means are so low, their program impacts are often significantly more positive than the 2017 cohort; there was no room for the decline that occurred in 2017. This echoes the larger targeting point of the paper: that there are bigger declines for youth at higher risk of the relevant outcomes.

Tables A14 and A15 show a similar pattern by gender. The control means for arrests in both cities reflect boys' disproportionate involvement in the criminal justice system. Boys in the sample are also considerably more likely to have received substance abuse treatment (1.8 versus 0.2 percent) and less likely to persist in school (89 versus 93 percent in Philadelphia, and 94 versus 96 percent in Chicago). Point estimates on arrests are generally bigger for boys, though in Chicago the declines for women are still statistically significant on their own.

In Philadelphia, males have a proportionally huge and statistically significant decline in substance abuse treatment (a 1.1 percentage point ITT decline relative to a control mean of 1.8 percent). This contributes to an overall drop in the receipt of behavioral health services for boys, 2.5 percentage points, or a 22 percent decline ($p = 0.057$). Males also show the only significant adverse effect, a 0.9 percentage point increase in parenthood, which is almost a tripling relative to the control mean. While it is certaintly possible that the program increases confidence and income in a way that increases risky sexual activity, it is also true that there is more of a margin for increases in reporting of childbirth for fathers than for monthers (who have a negative but not significant point estimate).

Table A16 shows Chicago results separated by whether someone had a prior arrest. As expected, the control means on arrests are much higher for those with a prior arrest record. The point estimates also tend to be much larger, though they are only statistically differentiable from each other for other arrests. Both groups have significant declines in arrests, so it's not that lower-risk groups do not respond at all. But the groups move proportional to their baseline, so that declines tend to be bigger for the more criminally-involved group.

Table A17 shows Philadelphia results by race. Since some of the racial and ethnic groups in the data are too small to estimate separate effects for, the analysis just divides the sample into Black and non-Black, which includes Hispanic, white, Asian, and other. This is not because these groups necessarily all have the same treatment effect, but rather because there is not enough data to estimate separate effects. Incarceration declines are significantly bigger among Black youth, though the point estimates for arrests are suggestively larger for non-Blacks. The other significant difference is in the receipt of child protective services; Black youth experience a significant 0.9 percentage point (43 percent) decline in the prevalence of these services, with no change for non-Black youth.

The final interaction table, Table A18, shows treatment effects for those under 16 versus 16 and over. Significant effects are typically concentrated among the younger group, other than a decline in substance abuse for the older group. But differences between age groups are not significant.

F Risk-Responsiveness

Appendix Tables A19 and A20 show alternative versions of the endogenous stratification exercise using more expansive indices. Rather than limit each index to the outcomes where the SYEP has significant main effects, these tables use all the available socially costly outcomes. This avoids cherry-picking outcomes based on ex-post significance, but it limits power by combining outcomes that the intervention does not move with outcomes that are affected.

In Philadelphia, the index includes juvenile incarceration, juvenile justice services, the 4 types of arrests separately (excluding total since that is the sum of the rest), parenthood, an indicator for child protective services, and an indicator for using a housing shelter. In Chicago, it includes each of the 4 types of arrests separately. In Philadelphia, the pattern of results is quite similar, with the highest-risk groups having the largest point estimates. But as expected, the estimates are noisier. In Chicago, the overall impact on the index is no longer significant, and the point estimates flatten out across risk groups. This suggests that the outcomes not included in the original index (violent and property crimes) may have a somewhat different pattern of responses there. But the standard errors are big enough that this may also just be additional noise from including more skewed outcomes without detectable treatment effects; unlike in Philadelphia and the more limited Chicago index, the standard errors on the highest-risk group are over 5 times as large as the point estimate.

The main text's discussion of Figure 2 refers to this section for a list of which estimates are included in each panel. The details are as follows: Panel A includes Philadelphia coefficients for main effects on total and other arrests, juvenile incarceration, child protective services, and year 2 property crime arrests. For OSC+, it includes main effects on drug and other arrests, and year 3 arrests for all, property, and other crimes. Panel B adds the decline in violent-crime arrest and increase in years 2 and 3 for property crime arrests in OSC+ 2012; the decline in violent-crime arrests for OSC+ 2013; the decline in incarceration and mortality from NYC's SYEP; and the decline in violent and property crime arraignments in Boston (Davis and Heller, 2020; Modestino, 2019; Gelber et al., 2016). We note that Kessler et al. (2021) was released shortly before this paper. Because that study is still a working paper, and point estimates sometimes move around somewhat over the review process, we do not include those estimates in the graph; all estimates in these graphs have been through peer review. The pattern in their findings is, however, quite consistent with the one shown here.

The subgroups added in Panel C include behavioral health services, incarceration, total and drug arrests, parenthood, and substance abuse for boys in WorkReady; other arrests for girls in WorkReady; incarceration, total and other arrests, and child protection services for the 2017 cohort in WorkReady; incarceration, other arrests, child protection services, and behavioral health services for Black youth in WorkReady; total and property arrests for non-Black youth in WorkReady; total, drug, and other arrests for those under age 16 in WorkReady; substance abuse services for those 16 and older in WorkReady; drug arrests for women in OSC+ 2015; other arrests for those with a prior arrest and drug arrests for those with no prior arrest in OSC+ 2015; incarceration impacts for both age groups, for men, whites, and Blacks, for those not working, and those not living in an empowerment zone in NYC; mortality effects for men, Latinos, younger people, those not working, those living in an empowerment zone, and those observed through the 9-year follow-up in NYC; violent-crime arrests for in-school youth, males, females, those with prior arrests, and those without prior arrests in the pooled OSC+ 2012 and 2013 sample; and drug arrests for in-school youth and property-crime arrests for females in the pooled OSC+ 2012 and 2013 sample. The Boston study does not report control means by subgroup, so no subgroup effects are included from that study. Additionally, the Boston study does not report LATE estimates; the main graph backs out the implied LATE from the reported ITT and the reported take-up rate.

Figure A.1, showing the ITT instead of the LATE, reflects a similar pattern but is slightly more noisy. This suggests that some of the differences across groups is in the take-up rate. Although those differences may be informative for policy (e.g., if some types of youth are easier to enroll), some of the take-up differences are really due to differences in experimental designs across studies, i.e., encouragement designs versus studies that strongly enforced compliance. In the ITT version of these graphs, there is also some additional error in our calculation of the NYC subgroup estimates. This is because the original paper only reports point estimates from two-stage least squares. Here we back out intent-to-treat estimates by multiplying IV estimates by the first stage. But the paper only reports the overall first stage. If take-up varies by subgroup, these estimates will mis-estimate the subgroup ITTs.

References

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

Davis, Jonathan M.V. and Heller, Sara B. (Oct. 2020). "Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs". *The Review of Economics and Statistics* 102(4), pp. 664–677.

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

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

Kessler, Judd B., Tahamont, Sarah, Gelber, Alexander M., and Isen, Adam (2021). "The Effects of Youth Employment on Crime: Evidence from New York City Lotteries". *NBER Working Paper No. 28373*.

McClanahan, Wendy S., Sipe, Cynthia L., and Smith, Thomas J. (2004). "Enriching Summer Work: an Evaluation of the Summer Career Exploration Program". *Public/Private Ventures*.

Modestino, Alicia Sasser (2019). "How Do Summer Youth Employment Programs Improve Criminal Justice Outcomes, and for Whom?" *Journal of Public Policy Analysis and Management*.

Morgan, Kari and Rubin, Donald B. (2012). "Rerandomization to improve covariate balance in experiments". *Annals of Statistics* 40(2).2, pp. 1263–1282.

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

G Tables

Table A1: Program Impacts in First Year After Random Assignment, No Baseline Covariates

	ITT	CM	LATE	CCM
	WorkReady			
Any Juvenile Incarceration	-0.006**	0.014	-0.019**	0.023
	(0.003)		(0.009)	
Any Receipt of Juvenile Justice Services	-0.004	0.013	-0.012	0.023
	(0.003)		(0.010)	
Total Number of Arrests	-0.012**	0.028	-0.034**	0.050
	(0.005)		(0.015)	
Number of Violent Arrests	-0.003	0.011	-0.009	0.019
	(0.004)		(0.010)	
Number of Property Arrests	-0.002	0.008	-0.007	0.013
	(0.003)		(0.008)	
Number of Drug Arrests	-0.003*	0.004	-0.008*	0.008
	(0.002)		(0.005)	
Number of Other Arrests	-0.003**	0.004	-0.010**	0.009
	(0.001)		(0.004)	
Graduated or Still Enrolled in School	-0.001	0.913	-0.002	0.937
	(0.010)		(0.026)	
	OSC+			
Total Number of Arrests	-0.030*	0.176	-0.114*	0.193
	(0.017)		(0.065)	
Number of Violent Arrests	0.004	0.037	0.013	0.019
	(0.006)		(0.023)	
Number of Property Arrests	0.000	0.020	0.001	0.020
	(0.005)		(0.018)	
Number of Drug Arrests	-0.014**	0.034	-0.052**	0.058
	(0.006)		(0.024)	
Number of Other Arrests	-0.020**	0.084	-0.076**	0.095
	(0.010)		(0.038)	
Graduated or Still Enrolled in School	0.004	0.951	0.015	0.961
	(0.007)		(0.025)	

Note: WorkReady N=4497, OSC+ N=5405. Graduated/still enrolled excludes pre-program graduates and those unmatched to education records; WorkReady N=3858, OSC+ N=4077. Table shows estimated intent-to-treat (ITT) and local average treatment effects (LATE), controlling for randomization block. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses, clustered by person for WorkReady, where the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A2: WorkReady Family and Health Impacts in First Year After Random Assignment, No Baseline Covariates

	ITT	CM	LATE	CCM
Becomes A Parent	-0.001	0.011	-0.002	0.004
	(0.003)		(0.009)	
Any Receipt of Child Protective Services	-0.008**	0.018	-0.023**	0.020
	(0.003)		(0.010)	
Any Stay in a Shelter	0.001	0.001	0.004	0.000
	(0.002)		(0.004)	
Any Receipt of Behavioral Health Services	-0.015	0.113	-0.045	0.119
	(0.010)		(0.029)	
Any Receipt of Substance Abuse Services	-0.003	0.008	-0.009	0.010
	(0.003)		(0.008)	
Any Receipt of Mental Health Services	-0.014	0.110	-0.041	0.115
	(0.010)		(0.029)	

Note: N=4497. Table shows estimated intent-to-treat (ITT) and local average treatment effects (LATE), controlling for randomization block. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. Becoming a parent is measured across all available outcome years; other outcomes in first post-randomization year only. Robust standard errors in parentheses, clustered by person as the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A3: Intent-to-Treat Results with Alternative Functional Forms, Probit or Poisson

	WorkReady		OSC+	
	AME	CM	AME	CM
Any Juvenile Incarceration	-0.008*	0.019		
	(0.005)			
Any Receipt of Juvenile Justice Services	-0.004	0.017		
	(0.004)			
Total Number of Arrests	-0.011*	0.028	-0.028*	0.176
	(0.006)		(0.016)	
Number of Violent Arrests	-0.003	0.011	0.004	0.037
	(0.004)		(0.006)	
Number of Property Arrests	-0.003	0.011	0.000	0.020
	(0.003)		(0.005)	
Number of Drug Arrests	-0.013	0.020	-0.013**	0.034
	(0.009)		(0.006)	
Number of Other Arrests	-0.006**	0.007	-0.019*	0.084
	(0.003)		(0.010)	
Graduated or Still Enrolled in School	-0.001	0.911	0.000	0.951
	(0.010)		(0.006)	
Parenthood	0.000	0.013		
	(0.004)			
Receipt of Child Protective Services	-0.010**	0.022		
	(0.005)			
Any Stay in a Shelter	0.003	0.003		
	(0.003)			
Receipt of Behavioral Health Services	-0.012	0.114		
	(0.009)			
Substance Abuse Services	-0.002	0.011		
	(0.003)			
Mental Health Services	-0.011	0.111		
	(0.008)			

Note: WorkReady N=4497, OSC+ N=5405. Graduated/still enrolled excludes pre-program graduates and those unmatched to education records; WorkReady N=3858, OSC+ N=4077. AME indicates average marginal effects. Total, violent, and property crimes use Poisson regression; all others only take on values 0 and 1, so use probit. CM is control mean. Robust standard errors in parentheses, clustered by person for WorkReady where the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A4: Multiple Testing by Family, WorkReady

	Standard P-Value	Randomization Inference P-Value	FWER Adjustment	FDR Adjustment
Any Crime				
Any Juvenile Incarceration	0.08	0.10	0.15	0.12
Any Receipt of Juvenile Justice Services	0.47	0.49	0.48	0.47
Total Number of Arrests	0.04	0.05	0.11	0.12
Type of Crime				
Number of Violent Arrests	0.53	0.56	0.73	0.53
Number of Property Arrests	0.48	0.47	0.73	0.53
Number of Drug Arrests	0.10	0.11	0.29	0.21
Number of Other Arrests	0.01	0.01	0.05	0.04
Graduated or Still Enrolled in School	0.50	0.48		
Family				
Becomes a Parent	0.81	0.79	0.81	0.81
Any Receipt of Child Protective Services	0.09	0.12	0.28	0.28
Any Stay in a Shelter	0.19	0.16	0.36	0.29
Health				
Any Receipt of Substance Abuse Services	0.30	0.27	0.43	0.31
Any Receipt of Mental Health Services	0.26	0.27	0.43	0.31

Note: N=4497. Table shows the adjusted p-values controlling for the family-wise error rate (FWER) and q-values showing the smallest false discovery rate (FDR) under which each null can be rejected. Graduated/still enrolled excludes pre-program graduates and those unmatched to education records; N=3858.

Table A5: Multiple Testing by Family, OSC+

	Standard P-Value	Randomization Inference P-Value	FWER Adjustment	FDR Adjustment
Total Number of Arrests	0.13	0.12		
Type of Crime				
Number of Violent Arrests	0.35	0.34	0.58	0.47
Number of Property Arrests	0.93	0.92	0.93	0.93
Number of Drug Arrests	0.04	0.04	0.15	0.13
Number of Other Arrests	0.07	0.07	0.19	0.13
Graduated or Still Enrolled in School	0.91	0.93		

Note: N=5405. Table shows the adjusted p-values controlling for the family-wise error rate (FWER) and q-values showing the smallest false discovery rate (FDR) under which each null can be rejected. Graduated/still enrolled excludes pre-program graduates and those unmatched to education records; N=4077.

Table A6: Intent to Treat on Education Outcomes in the First Year After Randomization with Imputations, WorkReady

	Grade Point Average	Any Suspensions	Days Absent
Panel A. Non-Missing Only			
Intent to Treat	-0.033	0.011	-0.035
	(0.036)	(0.013)	(0.733)
CM	2.242	0.108	15.3
N	2021	2166	2166
Panel B. Imputed Mean (N=4211)			
Intent to Treat	-0.009	0.004	-0.320
	(0.018)	(0.007)	(0.380)
CM	2.244	0.108	15.2
Panel C. Low-Performing Imputation (N=4211)			
Intent to Treat	-0.023	0.005	0.021
	(0.027)	(0.011)	(0.532)
CM	1.081	0.539	34.5
Panel D. High-Performing Imputation (N=4211)			
Intent to Treat	-0.007	0.004	-0.084
	(0.026)	(0.007)	(0.415)
CM	3.049	0.056	7.9
Panel E. Low-Performing Imputation for Non-Transfers (N=4211)			
Intent to Treat	-0.014	0.004	-0.227
	(0.028)	(0.011)	(0.535)
CM	1.130	0.528	34.0
Panel F. High-Performing Imputation for Non-Transfers (N=4211)			
Intent to Treat	-0.014	0.005	-0.196
	(0.026)	(0.007)	(0.409)
CM	3.023	0.057	8.1

Note: Exludes pre-program graduates. Table shows estimated intent-to-treat effects, controlling for baseline covariates and randomization block, for various imputations. Non-Missing shows the treatment-control difference for non-missing data only. Imputed Mean imputes the mean outcome variable by group (treatment and control) and study year for all missing data. Low- and High-Performing Imputation assume all missing data is missing for either the most low- or high-performing students respectively. Low- and High- Performing for Non-Transfers assume data are missing completely at random (conditional on year and group) for those who are marked as transferring out of the district in the school records, then they assign all-low or all-high values for the remaining missingness. CM is control mean. Robust standard errors in parentheses, clustered by person as the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A7: Local Average Treatment Effect on Education Outcomes in the First Year After Randomization with Imputations, WorkReady

	Grade Point Average	Any Suspensions	Days Absent
Panel A. Non-Missing Only			
Local Average Treatment Effect	-0.100	0.034	-0.112
	(0.110)	(0.040)	(2.265)
CCM	2.422	0.066	12.6
N	2021	2166	2166
Panel B. Imputed Mean (N=4211)			
Local Average Treatment Effect	-0.026	0.012	-0.926
	(0.053)	(0.019)	(1.089)
CCM	2.332	0.091	13.9
Panel C. Low-Performing Imputation (N=4211)			
Local Average Treatment Effect	-0.066	0.016	0.059
	(0.078)	(0.032)	(1.524)
CCM	1.134	0.552	34.4
Panel D. High-Performing Imputation (N=4211)			
Local Average Treatment Effect	-0.019	0.010	-0.242
	(0.074)	(0.019)	(1.188)
CCM	3.135	0.039	6.2
Panel E. Low-Performing Imputation for Non-Transfers (N=4211)			
Local Average Treatment Effect	-0.040	0.011	-0.656
	(0.080)	(0.032)	(1.532)
CCM	1.163	0.542	34.3
Panel F. High-Performing Imputation for Non-Transfers (N=4211)			
Local Average Treatment Effect	-0.040	0.013	-0.568
	(0.074)	(0.019)	(1.172)
CCM	3.122	0.038	6.6

Note: Exludes pre-program graduates. Table shows estimated local average treatment effects (LATE), controlling for baseline covariates and randomization block, for various imputations. Non-Missing shows the treatment-control difference for non-missing data only. Imputed Mean imputes the mean outcome variable by group (treatment and control) and study year for all missing data. Low- and High-Performing Imputation assume all missing data is missing for either the most low- or high-performing students respectively. Low- and High- Performing for Non-Transfers assume data are missing completely at random (conditional on year and group) for those who are marked as transferring out of the district in the school records, then they assign all-low or all-high values for the remaining missingness. CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses, clustered by person as the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A8: Intent to Treat on Education Outcomes in the First Year After Randomization with Imputations, OSC+

	Grade Point Average	Any Suspensions	Days Absent
Panel A. Non-Missing Only			
Intent to Treat	-0.025	0.024**	0.159
	(0.024)	(0.012)	(0.559)
CM	2.394	0.153	25.1
N	3461	3851	3851
Panel B. Imputed Mean (N=4102)			
Intent to Treat	-0.012	0.024**	0.083
	(0.021)	(0.011)	(0.536)
CM	2.394	0.153	25.1
Panel C. Low-Performing Imputation (N=4102)			
Intent to Treat	-0.016	0.030**	0.436
	(0.029)	(0.012)	(0.597)
CM	2.014	0.204	27.8
Panel D. High-Performing Imputation (N=4102)			
Intent to Treat	-0.005	0.021*	-0.158
	(0.027)	(0.011)	(0.572)
CM	2.604	0.144	23.6
Panel E. Low-Performing Imputation for Non-Transfers (N=4102)			
Intent to Treat	-0.017	0.029**	0.354
	(0.027)	(0.012)	(0.566)
CM	2.134	0.178	26.4
Panel F. High-Performing Imputation for Non-Transfers (N=4102)			
Intent to Treat	-0.006	0.023**	-0.091
	(0.025)	(0.011)	(0.554)
CM	2.538	0.149	24.4

Note: Exludes pre-program graduates. Table shows estimated intent-to-treat effects, controlling for baseline covariates and randomization block, for various imputations. Non-Missing shows the treatment-control difference for non-missing data only. Imputed Mean imputes the mean outcome variable by group (treatment and control) and study year for all missing data. Low- and High-Performing Imputation assume all missing data is missing for either the most low- or high-performing students respectively. Low- and High- Performing for Non-Transfers assume data are missing completely at random (conditional on year and group) for those who are marked as transferring out of the district in the school records, then they assign all-low or all-high values for the remaining missingness. CM is control mean. Robust standard errors in parentheses. *p<0.1, **p<0.05, ***p<0.01

Table A9: Local Average Treatment Effect on Education Outcomes in the First Year After Randomization with Imputations, OSC+

	Grade Point Average	Any Suspensions	Days Absent
Panel A. Non-Missing Only			
Local Average Treatment Effect	-0.095	0.087**	0.581
	(0.089)	(0.043)	(2.035)
CCM	2.422	0.159	25.6
N	3461	3851	3851
Panel B. Imputed Mean (N=4102)			
Local Average Treatment Effect	-0.043	0.092**	0.312
	(0.079)	(0.042)	(2.002)
CCM	2.376	0.152	25.8
Panel C. Low-Performing Imputation (N=4102)			
Local Average Treatment Effect	-0.060	0.112**	1.635
	(0.107)	(0.047)	(2.234)
CCM	2.064	0.160	26.1
Panel D. High-Performing Imputation (N=4102)			
Local Average Treatment Effect	-0.020	0.080*	-0.591
	(0.099)	(0.042)	(2.136)
CCM	2.540	0.156	25.8
Panel E. Low-Performing Imputation for Non-Transfers (N=4102)			
Local Average Treatment Effect	-0.063	0.108**	1.327
	(0.101)	(0.044)	(2.117)
CCM	2.143	0.147	25.5
Panel F. High-Performing Imputation for Non-Transfers (N=4102)			
Local Average Treatment Effect	-0.023	0.084**	-0.342
	(0.094)	(0.042)	(2.070)
CCM	2.500	0.156	26.1

Note: Exludes pre-program graduates. Table shows estimated local average treatment effects (LATE), controlling for baseline covariates and randomization block, for various imputations. Non-Missing shows the treatment-control difference for non-missing data only. Imputed Mean imputes the mean outcome variable by group (treatment and control) and study year for all missing data. Low- and High-Performing Imputation assume all missing data is missing for either the most low- or high-performing students respectively. Low- and High- Performing for Non-Transfers assume data are missing completely at random (conditional on year and group) for those who are marked as transferring out of the district in the school records, then they assign all-low or all-high values for the remaining missingness. CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses. *p<0.1, **p<0.05, ***p<0.01

Table A10: Education Outcomes in the First Year After Randomization by Treatment Arm, OSC+

	ITT	CM	ITT	CM	Test of
	Job Only		Job and Mentor		Difference
Graduated or Still Enrolled in School (Year One)	-0.005	0.947	0.004	0.954	0.296
	(0.008)		(0.008)		
Days Absent	0.488	24.7	-0.162	25.5	0.439
	(0.690)		(0.708)		
Grade Point Average	-0.039	2.404	-0.012	2.385	0.430
	(0.030)		(0.029)		
Any Suspensions	0.027*	0.143	0.020	0.163	0.697
	(0.015)		(0.014)		
Graduated or Still Enrolled in School (Year Two)	-0.026**	0.895	-0.006	0.900	0.138
	(0.012)		(0.011)		

Note: Excludes pre-program graduates and those unmatched to education records; N=4407. Regressions use non-missing data only, N=3851 for days absent and suspensions and N=3461 for GPA. Table shows estimated intent-to-treat (ITT), controlling for baseline covariates and randomization block. CM is control mean. The Test of Difference column shows the p-value from the test that the treatment coefficients for each treatment arm are equal. Robust standard errors in parentheses. *p<0.1, **p<0.05, ***p<0.01

Table A11: WorkReady Descriptive Statistics and Baseline Balance by Cohort

	2017			2018		
	Control Mean	Treatment Mean	Test of Difference	Control Mean	Treatment Mean	Test of Difference
N	2,056	1,336	3,392	655	450	1,105
Demographics						
Age	15.7	15.8	0.98	15.4	15.4	0.80
Male	0.39	0.41	0.47	0.39	0.38	0.75
Black	0.77	0.76	0.73	0.83	0.80	0.24
Hispanic	0.13	0.13	0.80	0.07	0.10	0.13
Other Race	0.05	0.05	0.88	0.08	0.10	0.35
Is a Parent	0.02	0.02	0.52	0.01	0.00	0.49
Contact with Justice System						
Ever Incarcerated	0.027	0.024	0.56	0.008	0.000	0.025
Ever Received Juvenile Justice Services	0.028	0.022	0.37	0.012	0.011	0.866
Ever Arrested	0.053	0.041	0.10	0.024	0.033	0.392
Total Number of Prior Arrests	0.068	0.055	0.23	0.037	0.033	0.813
Violent	0.036	0.020	0.01	0.021	0.016	0.591
Property	0.021	0.021	0.68	0.005	0.011	0.244
Drug	0.004	0.003	0.98	0.000	0.000	.
Other	0.006	0.011	0.19	0.011	0.007	0.470
Receipt of Social Services						
Ever Received Child Protective Services	0.16	0.14	0.15	0.13	0.10	0.29
Ever Stayed in a Shelter	0.05	0.04	0.08	0.05	0.02	0.02
Any Behavioral Health Service	0.28	0.27	0.69	0.25	0.21	0.16
Any Substance Abuse Services	0.02	0.02	0.76	0.01	0.00	0.11
Any Mental Health Services	0.28	0.26	0.70	0.24	0.21	0.20
Educational Characteristics						
Enrolled in School	0.87	0.85	0.47	0.88	0.90	0.20
Graduated	0.07	0.09	0.29	0.05	0.05	0.66
Grade (self-reported)	9.7	9.8	0.81	9.0	9.0	0.85
Days Absent (if enrolled)	20.3	20.1	0.83	10.8	9.5	0.17
Grade Point Average (if non-missing)	2.38	2.44	0.24	2.50	2.50	0.96
Ever Suspended (if enrolled)	0.17	0.14	0.14	0.11	0.14	0.37
P-value on F-test of treatment-control comparison for all non-behavioral health baseline characteristics			0.52			0.28
P-value on F-test of treatment-control comparison for all behavioral health baseline characteristics			0.96			0.28

Note: Grade information is self-reported, N=4482; school absence (N=2336), GPA (N=2228), and suspension (N=2337) measures reflect only youth enrolled in public, non-charter schools in the School District of Philadelphia for whom data was non-missing.One WorkReady participant was missing race information. Enrolled and graduated are mutually exclusive. The difference column shows the p-value from the test that treatment and control means are equal, adjusting for randomization block. Behavioral health services, including substance abuse and mental health services, are held in a separate data set to maintain confidentiality of HIPAA-covered data and are thus a separate F-test from other baseline characteristics.

Table A12: Outcomes in the Second and Third Year After Randomization, by Year

	WorkReady Year Two				OSC+ Year Two				OSC+ Year Three			
	ITT	CM	LATE	CCM	ITT	CM	LATE	CCM	ITT	CM	LATE	CCM
Any Juvenile Incarceration	-0.006	0.017	-0.023	0.022								
	(0.004)		(0.014)									
Receipt of Juvenile Justice Services	0.001	0.015	0.003	0.004								
	(0.004)		(0.015)									
Total Number of Arrests	-0.004	0.027	-0.016	0.021	-0.011	0.123	-0.043	0.133	-0.022*	0.110	-0.082*	0.165
	(0.006)		(0.024)		(0.013)		(0.049)		(0.011)		(0.043)	
Number of Violent Arrests	0.000	0.012	-0.001	0.000	-0.005	0.025	-0.017	0.029	0.004	0.016	0.015	0.019
	(0.004)		(0.015)		(0.005)		(0.017)		(0.004)		(0.014)	
Number of Property Arrests	-0.006**	0.009	-0.022**	0.021	-0.003	0.018	-0.011	0.028	-0.006*	0.014	-0.023*	0.031
	(0.003)		(0.011)		(0.004)		(0.014)		(0.004)		(0.013)	
Number of Drug Arrests	0.003	0.003	0.013	0.000	0.002	0.013	0.006	0.016	-0.007	0.018	-0.025	0.025
	(0.003)		(0.012)		(0.004)		(0.016)		(0.004)		(0.016)	
Number of Other Arrests	-0.002	0.003	-0.006	0.007	-0.005	0.068	-0.020	0.060	-0.013*	0.062	-0.050*	0.089
	(0.002)		(0.007)		(0.009)		(0.034)		(0.008)		(0.029)	
Graduated or Still Enrolled in School	0.006	0.902	0.023	0.936	-0.016*	0.897	-0.060*	0.971				
	(0.011)		(0.039)		(0.009)		(0.034)					
Receipt of Child Protective Services	-0.003	0.012	-0.011	0.010								
	(0.003)		(0.013)									
Any Stay in a Shelter	0.000	0.002	-0.001	0.002								
	(0.002)		(0.008)									
Receipt of Behavioral Health Services	0.001	0.097	0.003	0.061								
	(0.010)		(0.039)									
Substance Abuse Services	0.004	0.009	0.015	0.000								
	(0.004)		(0.015)									
Mental Health Services	-0.002	0.093	-0.008	0.067								
	(0.010)		(0.038)									

Note: Only 2017 WorkReady cohort observed for second year of data, N=3,392. OSC+ N=5,405. Enrolled/graduated excludes pre-program graduates and those unmatched to education records; WorkReady N=2869; OSC+ N=4077. Table shows estimated intent-to-treat (ITT) and local average treatment effects (LATE), controlling for baseline covariates and randomization block. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses, clustered by person for WorkReady where the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A13: WorkReady Outcomes in the First Year After Randomization, by Cohort

	Intent-to-Treat					Local Average Treatment Effect				
	2018	2017	2018 CM	2017 CM	Test of Difference	2018	2017	2018 CCM	2017 CCM	Test of Difference
Juvenile Incarceration	0.001	-0.007**	0.008	0.017	0.155	0.002	-0.029**	0.009	0.028	0.054
	(0.005)	(0.004)				(0.008)	(0.014)			
Receipt of Juvenile Justice Services	0.005	-0.005	0.006	0.015	0.108	0.009	-0.019	0.006	0.028	0.091
	(0.005)	(0.004)				(0.009)	(0.014)			
Total Number of Arrests	0.001	-0.014**	0.012	0.033	0.106	0.003	-0.056**	0.013	0.072	0.030
	(0.008)	(0.006)				(0.013)	(0.024)			
Number of Violent Arrests	0.001	-0.003	0.008	0.013	0.614	0.001	-0.012	0.014	0.018	0.475
	(0.007)	(0.004)				(0.011)	(0.016)			
Number of Property Arrests	0.002	-0.003	0.003	0.010	0.282	0.004	-0.013	0.000	0.024	0.242
	(0.004)	(0.003)				(0.007)	(0.013)			
Number of Drug Arrests	-0.001	-0.003	0.002	0.004	0.440	-0.002	-0.012	0.002	0.012	0.236
	(0.001)	(0.004)				(0.002)	(0.008)			
Number of Other Arrests	0	-0.005***	0.000	0.006	0.014	-0.001	-0.019***	0.001	0.018	0.011
	(0.000)	(0.002)				(0.001)	(0.007)			
Graduated or Still in School	0.005	-0.008	0.945	0.944	0.374	0.008	-0.031	0.956	0.983	0.292
	(0.012)	(0.008)				(0.020)	(0.031)			
Becomes a Parent	0.001	0.001	0.000	0.015	0.947	0.002	0.003	0.000	0.001	0.952
	(0.001)	(0.004)				(0.002)	(0.017)			
Receipt of Child Protective Services	0.001	-0.008*	0.006	0.022	0.163	0.001	-0.03*	0.002	0.022	0.084
	(0.004)	(0.004)				(0.007)	(0.016)			
Any Stay in a Shelter	0.003	0.002	0.000	0.002	0.669	0.005	0.007	0.000	0.000	0.804
	(0.002)	(0.002)				(0.004)	(0.007)			
Receipt of Behavioral Health Services	0	-0.015	0.110	0.113	0.453	0	-0.057	0.104	0.110	0.252
	(0.017)	(0.010)				(0.031)	(0.039)			
Substance Abuse Services	0.001	-0.004	0.003	0.010	0.261	0.002	-0.016	0.002	0.014	0.211
	(0.003)	(0.003)				(0.006)	(0.013)			
Mental Health Services	-0.001	-0.013	0.110	0.109	0.534	-0.001	-0.05	0.106	0.103	0.327
	(0.017)	(0.010)				(0.031)	(0.039)			

Note: 2018 N=1780, 2017 N=2717. Enrolled/graduated excludes pre-program graduates and those unmatched to education records; N=3858. Table shows estimated intent-to-treat (ITT) and local average treatment effects (LATE), controlling for baseline covariates and randomization block. Coefficients estimated in single regression with interactions; linear combination of estimates and standard errors displayed to show net effect for each group. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses, clustered by person as the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A14: WorkReady Outcomes in the First Year After Randomization, by Gender

	Intent-to-Treat					Local Average Treatment Effect				
	Male	Female	Male CM	Female CM	Test of Difference	Male	Female	Male CCM	Female CCM	Test of Difference
Juvenile Incarceration	-0.011*	-0.001	0.029	0.005	0.123	-0.038*	-0.003	0.041	0.008	0.105
	(0.006)	(0.003)				(0.021)	(0.007)			
Receipt of Juvenile Justice Services	-0.006	0	0.025	0.005	0.320	-0.02	0.001	0.037	0.008	0.307
	(0.006)	(0.003)				(0.020)	(0.008)			
Total Number of Arrests	-0.019*	-0.005	0.049	0.015	0.201	-0.063*	-0.013	0.078	0.028	0.162
	(0.010)	(0.005)				(0.035)	(0.012)			
Number of Violent Arrests	-0.003	-0.002	0.019	0.007	0.838	-0.01	-0.004	0.026	0.011	0.801
	(0.007)	(0.003)				(0.023)	(0.008)			
Number of Property Arrests	-0.005	0	0.013	0.005	0.422	-0.016	0	0.017	0.009	0.396
	(0.005)	(0.003)				(0.017)	(0.008)			
Number of Drug Arrests	-0.006*	0	0.009	0.000	0.075	-0.021*	0	0.021	0.000	0.074
	(0.004)	(0.003)				(0.012)	(0.002)			
Number of Other Arrests	-0.005	-0.003**	0.008	0.002	0.575	-0.016	-0.008**	0.015	0.008	0.461
	(0.003)	(0.001)				(0.010)	(0.004)			
Graduated or Still in School	-0.005	-0.004	0.936	0.950	0.932	-0.017	-0.011	0.968	0.973	0.884
	(0.012)	(0.008)				(0.036)	(0.021)			
Parenthood	0.009**	-0.005	0.003	0.017	0.028	0.029**	-0.012	0.000	0.014	0.021
	(0.004)	(0.005)				(0.013)	(0.012)			
Receipt of Child Protective Services	-0.007	-0.005	0.017	0.019	0.741	-0.023	-0.013	0.016	0.012	0.614
	(0.005)	(0.005)				(0.015)	(0.012)			
Any Stay in a Shelter	0	0.003	0.002	0.001	0.188	0	0.009*	0.000	0.000	0.226
	(0.002)	(0.002)				(0.006)	(0.006)			
Receipt of Behavioral Health Services	-0.025*	-0.002	0.116	0.110	0.171	-0.089*	-0.004	0.126	0.098	0.125
	(0.013)	(0.011)				(0.047)	(0.031)			
Substance Abuse Services	-0.011**	0.003	0.018	0.002	0.003	-0.041**	0.009	0.046	0.000	0.003
	(0.004)	(0.003)				(0.016)	(0.008)			
Mental Health Services	-0.018	-0.004	0.108	0.110	0.412	-0.065	-0.011	0.102	0.105	0.331
	(0.013)	(0.011)				(0.046)	(0.031)			

Note: Male N=1780, Female N=2717. Enrolled/graduated excludes pre-program graduates and those unmatched to education records; N=3858. Table shows estimated intent-to-treat (ITT) and local average treatment effects (LATE), controlling for baseline covariates and randomization block. Coefficients estimated in single regression with interactions; linear combination of estimates and standard errors displayed to show net effect for each group. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses, clustered by person as the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A15: OSC+ Outcomes in the First Year After Randomization, by Gender

	Intent-to-Treat					Local Average Treatment Effect				
	Male	Female	Male CM	Female CM	Test of Difference	Male	Female	Male CCM	Female CCM	Test of Difference
Total Number of Arrests	-0.037	-0.014	0.330	0.069	0.49	-0.137	-0.053	0.299	0.081	0.51
	(0.033)	(0.012)				(0.120)	(0.046)			
Number of Violent Arrests	0.01	0.003	0.058	0.023	0.60	0.035	0.01	0.018	0.009	0.61
	(0.012)	(0.006)				(0.043)	(0.022)			
Number of Property Arrests	0.006	-0.003	0.027	0.016	0.34	0.023	-0.013	0.023	0.017	0.33
	(0.009)	(0.005)				(0.033)	(0.019)			
Number of Drug Arrests	-0.023	-0.005**	0.076	0.005	0.21	-0.083	-0.02**	0.107	0.016	0.22
	(0.014)	(0.002)				(0.051)	(0.009)			
Number of Other Arrests	-0.03	-0.008	0.169	0.026	0.29	-0.112	-0.03	0.151	0.039	0.30
	(0.020)	(0.007)				(0.075)	(0.026)			
Graduated or Still in School	0.006	-0.006	0.944	0.956	0.35	0.024	-0.021	0.975	0.980	0.35
	(0.010)	(0.008)				(0.039)	(0.029)			

Note: Male N=2168, Female N=3237. Enrolled/graduated excludes pre-program graduates and those unmatched to education records; N=4077. Table shows estimated intent-to-treat (ITT) and local average treatment effects (LATE), controlling for baseline covariates and randomization block. Coefficients estimated in single regression with interactions; linear combination of estimates and standard errors displayed to show net effect for each group. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses. *p<0.1, **p<0.05, ***p<0.01

Table A16: OSC+ Outcomes in the First Year After Randomization, by Prior Arrest Status

	Intent-to-Treat					Local Average Treatment Effect				
	Arrested	Not Arrested	Arrested CM	Not Arrested CM	Test of Difference	Arrested	Not Arrested	Arrested CCM	Not Arrested CCM	Test of Difference
Total Number of Arrests	-0.091	-0.009	0.592	0.056	0.20	-0.35	-0.033	0.503	0.089	0.20
	(0.063)	(0.010)				(0.245)	(0.036)			
Number of Violent Arrests	-0.001	0.005	0.112	0.016	0.75	-0.004	0.02	0.012	0.019	0.76
	(0.020)	(0.005)				(0.077)	(0.018)			
Number of Property Arrests	0.013	-0.003	0.057	0.010	0.37	0.052	-0.012	0.036	0.014	0.37
	(0.018)	(0.003)				(0.071)	(0.012)			
Number of Drug Arrests	-0.036	-0.007**	0.123	0.008	0.25	-0.139	-0.025**	0.151	0.029	0.25
	(0.026)	(0.003)				(0.100)	(0.010)			
Number of Other Arrests	-0.067*	-0.004	0.299	0.023	0.10	-0.259*	-0.017	0.305	0.027	0.11
	(0.038)	(0.005)				(0.149)	(0.020)			
Graduated or Still in School	0.01	-0.004	0.906	0.964	0.45	0.04	-0.014	0.975	0.979	0.45
	(0.017)	(0.006)				(0.068)	(0.023)			

Note: N=5405. "Arrested" refers to youth who had any record of arrest prior to randomization, regardless of number of arrests; Arrested N=1210, Not Arrested N=4195. Enrolled/graduated excludes pre-program graduates and those unmatched to education records; N=4077. Table shows estimated intent-to-treat (ITT) and local average treatment effects (LATE), controlling for baseline covariates and randomization block. Coefficients estimated in single regression with interactions; linear combination of estimates and standard errors displayed to show net effect for each group. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses. *p<0.1, **p<0.05, ***p<0.01

Table A17: WorkReady Outcomes in the First Year After Randomization, by Race

	Intent-to-Treat					Local Average Treatment Effect				
	Black	Non-Black	Black CM	Non-Black CM	Test of Difference	Black	Non-Black	Black CCM	Non-Black CCM	Test of Difference
Juvenile Incarceration	-0.007**	0.002	0.017	0.003	0.089	-0.02**	0.007	0.024	0.001	0.112
	(0.003)	(0.004)				(0.010)	(0.015)			
Receipt of Juvenile Justice Services	-0.001	-0.007	0.013	0.010	0.295	-0.002	-0.024	0.016	0.026	0.258
	(0.004)	(0.005)				(0.010)	(0.017)			
Total Number of Arrests	-0.009	-0.016**	0.030	0.021	0.415	-0.025	-0.055**	0.045	0.051	0.293
	(0.006)	(0.007)				(0.018)	(0.023)			
Number of Violent Arrests	-0.001	-0.007	0.012	0.010	0.383	-0.003	-0.022	0.016	0.019	0.335
	(0.004)	(0.005)				(0.012)	(0.017)			
Number of Property Arrests	-0.001	-0.006*	0.009	0.007	0.225	-0.002	-0.022*	0.009	0.022	0.179
	(0.003)	(0.003)				(0.009)	(0.012)			
Number of Drug Arrests	-0.003	-0.002	0.004	0.002	0.780	-0.008	-0.007	0.008	0.007	0.928
	(0.002)	(0.004)				(0.005)	(0.006)			
Number of Other Arrests	-0.005**	-0.001	0.005	0.002	0.170	-0.013**	-0.004	0.012	0.004	0.238
	(0.002)	(0.002)				(0.005)	(0.005)			
Graduated or Still in School	-0.006	0.001	0.937	0.974	0.569	-0.017	0.004	0.963	1.000	0.606
	(0.008)	(0.010)				(0.022)	(0.034)			
Becomes a Parent	0.003	-0.006	0.011	0.014	0.163	0.008	-0.021	0.000	0.018	0.164
	(0.004)	(0.006)				(0.010)	(0.019)			
Receipt of Child Protective Services	-0.009**	0.009	0.021	0.009	0.025	-0.027***	0.028	0.025	0.000	0.036
	(0.004)	(0.007)				(0.010)	(0.024)			
Any Stay in a Shelter	0.002	0.003	0.002	0.000	0.765	0.005	0.009	0.000	0.000	0.672
	(0.000)	(0.003)				(0.005)	(0.009)			
Receipt of Behavioral Health Services	-0.016	0.008	0.118	0.094	0.219	-0.047*	0.029	0.116	0.069	0.263
	(0.010)	(0.017)				(0.028)	(0.062)			
Substance Abuse Services	-0.004	0.003	0.008	0.007	0.216	-0.013	0.011	0.014	0.000	0.257
	(0.003)	(0.006)				(0.008)	(0.021)			
Mental Health Services	-0.013	0.003	0.114	0.094	0.413	-0.038	0.01	0.107	0.088	0.471
	(0.010)	(0.017)				(0.028)	(0.062)			

Note: Black N=3513, Non-Black N=984. Enrolled/graduated excludes pre-program graduates and those unmatched to education records; N=3858. Table shows estimated intent-to-treat (ITT) and local average treatment effects (LATE), controlling for baseline covariates and randomization block. Coefficients estimated in single regression with interactions; linear combination of estimates and standard errors displayed to show net effect for each group. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses, clustered by person as the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A18: WorkReady Outcomes in the First Year After Randomization, by Age

	Intent-to-Treat					Local Average Treatment Effect				
	Under 16	Over 16	Under 16 CM	Over 16 CM	Test of Difference	Under 16	Over 16	Under 16 CCM	Over 16 CCM	Test of Difference
Juvenile Incarceration	-0.004	-0.006	0.013	0.016	0.648	-0.01	-0.021	0.015	0.025	0.505
	(0.004)	(0.004)				(0.010)	(0.014)			
Receipt of Juvenile Justice Services	0.001	-0.006	0.010	0.015	0.217	0.004	-0.019	0.015	0.021	0.186
	(0.004)	(0.004)				(0.011)	(0.014)			
Total Number of Arrests	-0.011*	-0.009	0.028	0.028	0.817	-0.03*	-0.031	0.054	0.036	0.964
	(0.007)	(0.007)				(0.018)	(0.023)			
Number of Violent Arrests	0.002	-0.006	0.009	0.014	0.234	0.005	-0.02	0.016	0.016	0.207
	(0.005)	(0.005)				(0.012)	(0.016)			
Number of Property Arrests	-0.004	0	0.010	0.007	0.368	-0.011	0.001	0.016	0.006	0.421
	(0.004)	(0.004)				(0.010)	(0.012)			
Number of Drug Arrests	-0.003**	-0.002	0.004	0.004	0.490	-0.009**	-0.005	0.009	0.005	0.657
	(0.002)	(0.004)				(0.004)	(0.008)			
Number of Other Arrests	-0.006***	-0.002	0.005	0.004	0.222	-0.014***	-0.007	0.012	0.008	0.368
	(0.002)	(0.002)				(0.005)	(0.007)			
Graduated or Still in School	-0.001	-0.008	0.943	0.946	0.646	-0.003	-0.024	0.958	0.987	0.591
	(0.010)	(0.010)				(0.024)	(0.030)			
Becomes a Parent	-0.001	0.002	0.004	0.018	0.578	-0.003	0.008	0.001	0.000	0.591
	(0.002)	(0.006)				(0.006)	(0.019)			
Receipt of Child Protective Services	-0.006	-0.004	0.024	0.012	0.760	-0.017	-0.015	0.015	0.011	0.910
	(0.005)	(0.004)				(0.014)	(0.013)			
Any Stay in a Shelter	0.001	0.003	0.001	0.002	0.522	0.003	0.01	0.000	0.000	0.451
	(0.000)	(0.003)				(0.004)	(0.009)			
Receipt of Behavioral Health Services	-0.012	-0.009	0.121	0.104	0.839	-0.033	-0.033	0.083	0.139	1.000
	(0.012)	(0.012)				(0.032)	(0.042)			
Substance Abuse Services	0.001	-0.006*	0.004	0.012	0.106	0.003	-0.023*	0.000	0.025	0.083
	(0.003)	(0.004)				(0.009)	(0.013)			
Mental Health Services	-0.012	-0.007	0.120	0.099	0.742	-0.032	-0.025	0.083	0.130	0.883
	(0.012)	(0.012)				(0.032)	(0.042)			

Note: Under-16 N=2223, Over-16 N=2274. Enrolled/graduated excludes pre-program graduates and those unmatched to education records; N=3858. Table shows estimated intent-to-treat (ITT) and local average treatment effects (LATE), controlling for baseline covariates and randomization block. Coefficients estimated in single regression with interactions; linear combination of estimates and standard errors displayed to show net effect for each group. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. Robust standard errors in parentheses, clustered by person as the same person can appear in both cohorts. *p<0.1, **p<0.05, ***p<0.01

Table A19: WorkReady Treatment Effect on Combined Index by Predicted Risk Level

Panel A: Intent to Treat Effect, Index				
Full Sample	-0.045*			
	(0.026)			
Panel B: Intent to Treat Effect by Predicted Risk Level				
Predicted Risk Level	Repeated Split Sample	Leave One Out	CM	First Stage
Low	-0.011	0.038	-0.184	0.444
	(0.020)	(0.026)		
Medium	-0.019	-0.069*	-0.075	0.280
	(0.024)	(0.040)		
High	-0.099	-0.115	0.243	0.288
	(0.063)	(0.072)		

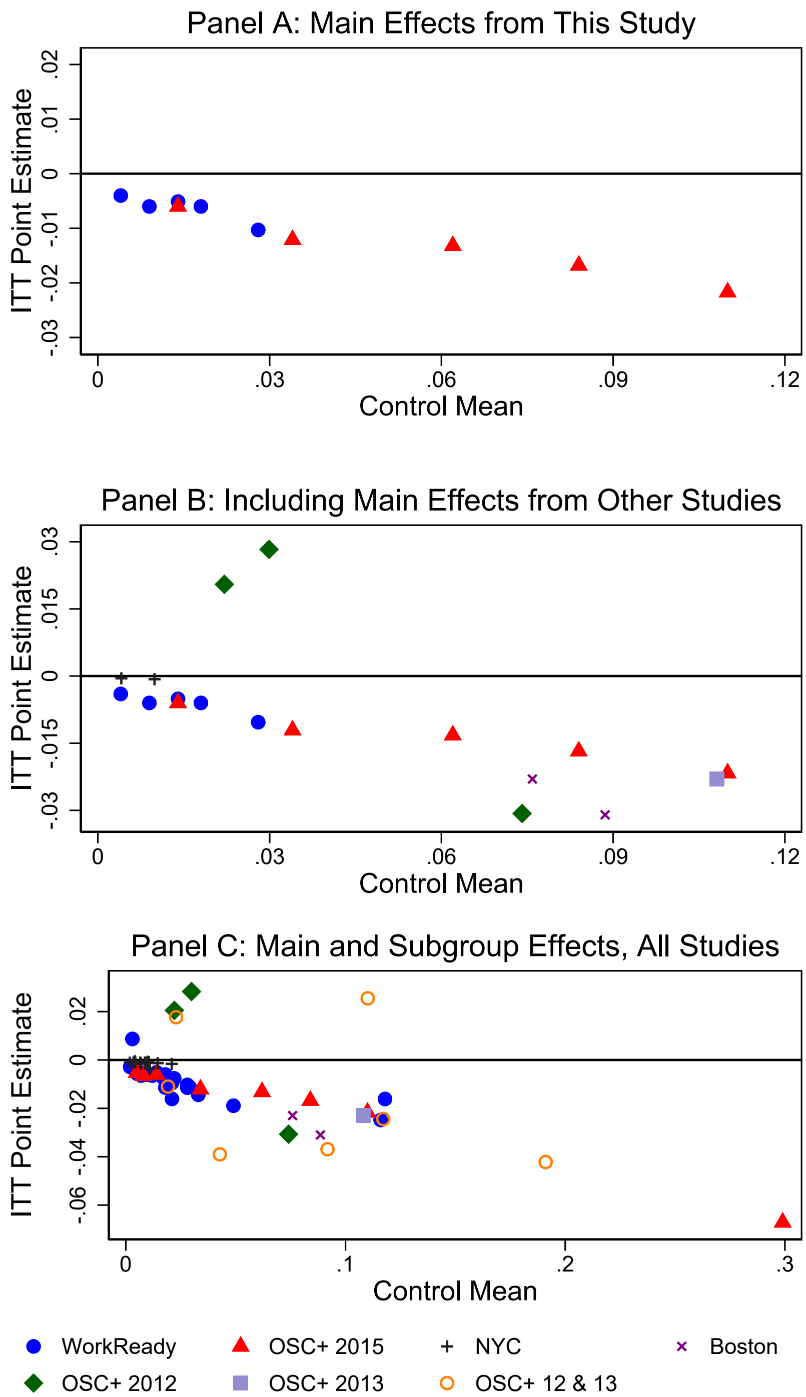
Note: N=4497. Estimation from the Abadie, Chingos, and West (2018) procedure. CM is the control mean for each group, assigned in the leave one out estimation. Table shows the effect of WorkReady on a standardized index of all the socially costly outcomes available in the WorkReady data: juvenile incarceration, juvenile justice services, the 4 types of arrests separately, excluding total since that is the sum of the rest, parenthood, an indicator for child protective services, and an indicator for using a housing shelter. Each variable is standardized on the control group, averaged together, then re-standardized so that the standard deviation of the index is 1. *p<0.1, **p<0.05, ***p<0.01

Table A20: OSC+ Treatment Effect on Combined Index by Predicted Risk Level

Panel A: Intent to Treat Effect, Index				
Full Sample	-0.023			
	(0.022)			
Panel B: Intent to Treat Effect by Predicted Risk Level				
Predicted Risk Level	Repeated Split Sample	Leave One Out	CM	First Stage
Low	-0.014	-0.022*	-0.208	0.263
	(0.010)	(0.013)		
Medium	-0.017	-0.009	-0.163	0.265
	(0.019)	(0.024)		
High	-0.012	-0.013	0.374	0.273
	(0.063)	(0.067)		

Note: N=5405. Estimation from the Abadie, Chingos, and West (2018) procedure. CM is the control mean for each group, assigned in the leave one out estimation. Table shows the effect of OSC+ on a standardized index of the 4 types of arrests separately, excluding total since that is the sum of the rest. Each variable is standardized on the control group, averaged together, then re-standardized so that the standard deviation of the index is 1. *p<0.1, **p<0.05, ***p<0.01

Figure A.1: Size of ITTs Relative to Control Means



Note: Point estimates and control means taken from this paper, Davis & Heller (2020), Gelber et al. (2016), and Modestino (2019). See text and Appendix F for details.