NBER WORKING PAPER SERIES

WHEN SCALE AND REPLICATION WORK: LEARNING FROM SUMMER YOUTH EMPLOYMENT EXPERIMENTS

Sara Heller

Working Paper 28705
http://www.nber.org/papers/w28705

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2021

This project was supported by Award No. 2016-R2-CX-0049, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, a State and Local Innovation Initiative grant (Philadelphia) and Social Policy Research Initiative grant (Chicago) from J-PAL North America, the Robert R. McCormick Foundation, and Project Development Grant Program funding from Poverty Solutions at the University of Michigan. Louise Geraghty, Brenda Mathias, Matt Repka, Misuzu Schexnider, and Lauren Shaw provided highly-accomplished project management; Kalen Flynn managed the Philadelphia qualitative data collection and analysis; Raquel Chavez, Kenny Hofmeister, Angela Hsu, Owen McCarthy, and Mary Clair Turner provided excellent research assistance; Marianne Bertrand provided invaluable support to the Chicago experiment, as did Greg Ridgeway for the Philadelphia study. The author thanks Jon Davis, Brian Jacob, and Basit Zafar for extremely helpful comments. We are grateful to the Chicago Department of Family and Support Services, the Philadelphia Youth Network, Inc., the Philadelphia mayor's office, and the University of Chicago Urban Labs for their partnership on these projects. We also thank the City of Philadelphia, the Philadelphia Police Department, the School District of Philadelphia, the Chicago Police Department, and the Chicago Public Schools for graciously allowing the use of their administrative data. Any further use of the data is subject to approval of each agency. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect those of these organizations. The studies are registered in the American Economic Association Registry under trial numbers 2451 (WorkReady) and 805 (OSC+). The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

When Scale and Replication Work: Learning from Summer Youth Employment Experiments
Sara Heller
NBER Working Paper No. 28705
April 2021
JEL No. I38,J08,K42

ABSTRACT

Because successful human capital interventions often fail to scale or replicate, public investment decisions require understanding how program size, context, and implementation shape program effects. This paper uses two new randomized controlled trials of summer youth employment programs in Chicago and Philadelphia to demonstrate how multiple experiments can help explain replicability and inform the expansion of promising approaches. Even when these programs grow or change models across contexts, participation consistently reduces criminal justice involvement. It may also decrease the need for child protective services and behavioral health treatment. Experimental variation in program model and local provider generates no detectable heterogeneity, suggesting that effects replicate partly because variability in implementation does not matter. There is, however, individual-level heterogeneity that explains differences in effect magnitudes across populations and informs optimal targeting; youth at higher risk of socially costly outcomes experience larger benefits. Identifying more interventions that combine this pattern of treatment heterogeneity with robust replicability could aid efforts to reduce social inequality efficiently.

Sara Heller
University of Michigan
Department of Economics
611 Tappan Street, Lorch Hall Room 238
Ann Arbor, MI 48109
and NBER
sbheller@umich.edu

An online appendix is available at http://www.nber.org/data-appendix/w28705

1 Introduction

As policymakers consider how to address economic and racial inequality, provide effective alternatives to policing, and find cost-effective ways to support residents' well-being, they must judge what evidence is promising enough to merit expanded investment. The existence of multiple randomized studies with similar findings is a common bar for considering an approach "evidenced-based." Research aggregators like Blueprints for Healthy Youth Development and the What Works Clearinghouse give their top ratings to programs with one or two high-quality randomized controlled trials (RCTs), with Blueprints calling programs in its top categories "ready for scale." Yet "ready for scale" and "scalable" are not the same thing. There are many examples of approaches that succeeded in one or two particular contexts, but had different effects when scaled up or moved elsewhere.[1] Key challenges include the increasing difficulty of finding talented staff as programs grow (Jepsen and Rivkin, 2009; Davis et al., 2017), the fact that bundled interventions make it hard to know which programmatic adaptations will matter, and, in a world with treatment heterogeneity, the fact that scale or setting may change who participates and thus how responsive participants are.

Summer youth employment programs (SYEPs) provide a useful example of how multiple sets of positive findings can establish the potential of an approach without answering core questions about how expanded investments would work. Experiments in Chicago, New York, and Boston have found generally similar patterns of SYEPs' effects: large declines in criminal justice involvement and violence, with little improvement in future employment on average (Heller, 2014; Davis and Heller, 2020; Modestino, 2019; Kessler et al., 2021; Gelber et al., 2016). Education impacts are more mixed, with most studies finding small or no improvements in high school or college outcomes (Schwartz et al., 2015; Leos-Urbel, 2014; Davis and Heller, 2020; Heller, 2014; Gelber et al., 2016), and one suggesting larger benefits (Modestino and Paulsen, 2019). But the Chicago and Boston studies focus on a relatively small subset of the cities' summer programs, with service providers selecting into the evaluation. So efforts to expand on these programs' success could generate changes in who is served or who is delivering the program, which might diminish their effectiveness.

[1] E.g., Perry Preschool, LIFE and CEO jobs programs, and Tennessee STAR among others.

NYC's SYEP is city-wide and at scale, but existing evaluations focus on how the program was implemented 15 years ago in a different economic and criminal justice context. So it is unclear whether changes in counterfactual opportunities affect how much of an impact SYEPs can have. To make decisions about where future investments are likely to succeed, policymakers require a better understanding of the determinants of replicability and scale.

This paper reports the results of two new randomized controlled trials, one in Chicago (n = 5,405) and one in Philadelphia (n = 4,497), designed to assess how SYEPs change when they scale, how implementation variation matters, and how heterogeneous populations respond to similar interventions across contexts. The first experiment tests a program called One Summer Chicago Plus (OSC+) in summer 2015. There were a few key changes from the 2012 and 2013 versions of the program studied elsewhere (Heller, 2014; Davis and Heller, 2020), including a tripling in the size of program that facilitates an analysis of how scale matters. The second experiment studies Philadelphia's WorkReady program in the summers of 2017 and 2018.[2] Unlike OSC+, which is a small subset of Chicago's varied summer programming, WorkReady is the umbrella program for the entire city's summer jobs program.

The paper does three things to assess and unpack questions of replicability and scale. First, it reports the two experiments' main effects to assess how consistent impacts are when approaches are scaled up or implemented differently in a previously unevaluated setting. In addition to using administrative data on outcomes that have been analyzed in other SYEP studies — criminal justice involvement and education — it also adds new measures of health and family well-being including child protective services, homeless shelter use, fertility, and receipt of substance abuse or mental health treatment. Second, it uses both experimental and non-experimental variation in program size, design, and delivery across local providers to assess how variation in these features shapes variation in treatment effects. Third, it leverages patterns of individual treatment heterogeneity across outcomes, subgroups, and studies to establish how effects may change as programs expand or policymakers make different

[2]The Philadelphia study registration includes a pre-analysis plan detailing primary and secondary hypotheses, as well as methods to address multiple testing concerns. The OSC+ study was pre-registered but without a pre-analysis plan, largely because the outcome definitions follow the prior studies of OSC+ exactly, limiting the scope for any potential data mining. We nonetheless perform similar multiple testing adjustments, described in the methods section below.

targeting decisions.[3]

The similarity of the main effects for programs that have recently grown and are operating at scale suggest SYEPs have unusual promise to replicate across contexts. They consistently reduce criminal justice contact in the first year after random assignment. Participating in WorkReady generated 3 fewer arrests per 100 participants (p = 0.043), a 65 percent decline relative to the control complier mean (CCM).[4] OSC+ showed a proportionally similar decline — a 52 percent drop from a higher baseline (LATE = -9 per 100 participants) — although the result is not quite statistically significant (p = 0.125). In both cities, these changes were driven by significant decreases in arrests for drug and other non-violent, non-property crime arrests ranging from 20 to 100 percent drops relative to baseline means. Arrests for violent crime did not significantly decrease, but the confidence intervals do not rule out the 30 – 40 percent declines found in other work. In Philadelphia, where we can measure juvenile incarceration, that too fell (LATE = -1.5 percentage points, an almost 80 percent decline, p = 0.09). For cohorts where enough time has passed to measure longer-term effects, there are also indications of declines in total, property, and other arrests during years 2 and 3. Neither city's participants show improvements in school engagement.

In Philadelphia, we can also measure other family and health outcomes. These outcomes might respond to program-generated improvements in self-efficacy, changes in peer networks, reduced exposure to crime, or just additional income that reduces the kinds of scarcity and stress that contribute to neglect and abuse. There are empirical indications of beneficial effects on some of these measures, including a 0.6 percentage point or 33 percent decline in the receipt of child protective services (p = 0.09). There was also a proportionally large decline in the receipt of mental health and substance abuse services (3.3 percentage points, or 31 percent), although this effect is not statistically different from 0 in the full sample, only in subgroups. There were no impacts on overall fertility or homeless shelter use, though both are quite infrequent in the sample. The imprecision of these results, especially after

[3]An upcoming companion project makes a fourth point, based on the observation that those who benefit the most are least likely to take-up the program on their own. Heller and Bhanot (in progress) use a separate "nudge" experiment, along with non-experimental variation in the level of administrative enrollment support, to demonstrate which strategies help the more responsive population overcome the higher barriers to participation they face.

[4]We focus on summarizing the local average treatment effects (LATEs) for compliers here to adjust for differences in take-up rates across studies; the text and tables also report intent-to-treat impacts.

adjustments for multiple testing, neccesitates some caution in interpretation. But they are promising enough to merit attention to these less-commonly measured aspects of well-being in future work.

Beyond main effects, the experimental design helps explain why SYEPs have proportionally similar crime effects across contexts by varying two factors that often generate replication challenges: who implements the programming and what program elements are included. Neither experiment is fully powered to test different explanations, in part because non-compliance resulted in first stages that are both around 0.3. Nonetheless, the results suggest that SYEPs might be particularly replicable and scalable because variation in implementation and program design does not generate variation in effects.

In this setting, the choice of local service provider affects the type of job placement, program staff, and in Philadelphia also the program model and type of professional development offered. To test whether this kind of implementation and staffing variation matter in practice, youth were randomly assigned to providers within randomization blocks in Chicago and a subset of Philadelphia. This design faciliates a well-identified test for excess variance in treatment effects across providers, requiring less power than estimating provider-specific treatment effects or testing for heterogeneity by provider characteristics. The treatment effects show no more significant variation across randomly-assigned provider than we would expect by chance, suggesting that the basic structure of the program, not the details of the delivery or the variation in program model, is most central for generating program effects.

To further explore how program structure matters, OSC+ included two treatment arms that randomly varied how youth spent their time in the program: either working all 25 hours per week with an adult supervisor, or working for 20 hours per week and participating in a civic leadership curriculum for 5 hours, with a more-involved adult mentor providing support and helping to handle barriers to employment. The difference in structure between OSC+ treatment arms — with and without the extra social supports of a mentor and civics curriculum — provides a hint that the additional supports SYEPs provide may matter over and above what a regular summer job would do. Significant improvements are concentrated in the richer program arm, with point estimates about twice as large as the job by itself. But the study lacks the power to fully differentiate the two versions.

The final part of the paper uses patterns of individual-level heterogeneity to establish how choices about whom to serve will shape program effects across contexts. We focus on hetereogeneity across different levels of baseline risk, presenting a framework to be concrete about how the shape of the relationship between counterfactual outcomes Y_0 and treatment effects can inform optimal targeting. Estimating this risk-responsiveness relationship may also draw out something deeper about the nature of heterogeneity; debates about the merits of prevention versus remediation are, in part, a question of whether some level of Y_0 is so high as to prevent responsiveness to the treatment.

We use two strategies to show that there is a consistently positive relationship between the risk of harmful outcomes a group faces and their responsivness. First, in each city, we generate an index of all the socially costly outcomes with significant treatment effects. We then use endogenous stratification to look at heterogeneity over that risk (Abadie et al., 2018). Point estimates suggest that the highest-risk group has a treatment effect between 13 and 26 times as large as the lowest-risk group. But as is often the case with subgroup analysis in a single study, standard errors make it hard to pin down the slope of the relationship between risk group and response.

The second strategy brings more information to the estimation of the risk-responsiveness relationship. We collect 62 statistically significant impact estimates from both the two experiments here, as well as four other peer-reviewed experimental studies of SYEPs. Across locations, outcomes, and subgroups, we plot each group's control mean for a socially-costly outcome — a measure of baseline risk — against its estimated LATE. The relationship between how common a costly outcome is in a subpopulation and the size of an SYEP's impact is strikingly linear; SYEPs generate bigger declines for populations where the outcome is more common. This pattern, along with the different eligibility criteria across cities, helps to explain why the effects of OSC+ are larger than WorkReady in absolute magnitude, though proportionally similar. And it provides guidance for how effects will change in new contexts or expanded programs: magnitudes are likely to scale with the risk level of the population served. More broadly, the fact that initially worse-off youth benefit the most suggests a virtuous complementarity between efficiency and equity in targeting decisions.

In practice, policymakers may want to prioritize program goals not measured here, such as

providing income transfers or developing labor market skills for a broader population. And if peer interaction is an important mechanism, major shifts in participant populations beyond what has already been tested could diminish treatment effects. Nonetheless, the study's findings suggest that shifting SYEPs' focus towards populations at elevated risk of crime and health problems could help maximize program benefits on those outcomes, and would likely do so across different scales, contexts, and program designs. The paper demonstrates how assessing the different elements required for a program to scale and replicate can provide crucial input into decisions about how and where to invest in human capital interventions.

2 Program Descriptions and Experimental Design

2.1 One Summer Chicago Plus

OSC+ is the 2015 version of the program that was evaluated in Heller (2014) and Davis and Heller (2020), run by Chicago's Department of Family and Support Services. The basic program structure was the same as reported elsewhere, with youth placed in non-profit, government, and private sector jobs, as well as participating in personal development programming, for 25 hours per week. But the program continued to develop over time in substantively important ways.[5] The initial 2012 cohort involved 700 program slots coordinated by 5 provider agencies in 2012. The 7-week 2015 program scaled up significantly, about tripling in size to 2,000 program slots delivered by 19 providers. To identify enough applicants to fill the new slots, recruiting occurred in 49 high schools located in high-violence neighborhoods, compared to 13 in 2012. All 16- to 21-year-old students in those schools were eligible to apply. Youth received a full week of work readiness training at the beginning of the program rather than just a day. And the program obtained a waiver against the city's new minimum wage increase, so youth were still paid $8.25 per hour (compared to the recent shift to a $10 minimum wage).

The largest substantive difference in the 2015 program was the removal of the original social-emotional learning curriculum. Instead, program operators tested two versions of the program to isolate the effects of some of the more expensive program elements. Half the

[5]It has also continued to change since 2015.

program youth had the opportunity to work in their jobs for the full 25 hours per week, with no separate adult mentor and no additional youth development curriculum. The other half worked for 20 hours per week, and on Fridays engaged in a Civic Leadership Foundation curriculum focused on civic leadership.[6] The initial purpose was to assess whether the additional costs associated with adult mentors and professional development were a key mechanism driving the previously-documented effects, though as discussed below, the reality of implementation makes this difficult to answer with certainty.[7]

A total of 5,405 youth applied to OSC+. The reseach team grouped them into geographic blocks to help minimize commute times. Within blocks, youth were randomly assigned to a provider and to the control group ($n = 2,911$), the job-only group ($n = 1,252$), or the job+mentor group ($n = 1,242$). This introduces random variation in provider, which we use to assess whether there is treatment heterogeneity by provider.

2.2 WorkReady

For over 15 years, PYN has offered a summer jobs program for youth ages 14 to 21 called WorkReady. During the study years, using a blend of government and private funding, PYN contracted with 50-60 local agencies to implement the six-week WorkReady summer program.[8] Youth accepted to the program were assigned to a local provider, which placed them in one of three program models: service learning to address a community problem, work experience with skill development and ongoing adult interaction, or an internship that included professional development and less intensive adult mentoring. All three models focused on developing "21st-Century Workforce Skills" and offered an hourly wage, but they varied in how like a private-sector summer job they were. Youth were not randomly assigned to the different program models, so the effects estimated here are the average of the models

[6]See https://www.civicleadershipfoundation.org/curriculum for details. This curriculum was quite different from prior OSC+ studies, where the programming focused on developing socio-emotional skills like self regulation, goal setting, and perspective-taking.

[7]This is both because of the compliance issues described below, and because program providers were quite resistant to removing the additional mentorship. They replaced the mentors with "adult supervisors" who purposefully provided less personalized and intensive support, but who were nonetheless available as needed. Site observations suggested that there was still a difference in the amount of adult support offered across treatment arms, but the difference in practice was not as stark as originally intended.

[8]Both the number of providers and the program models have continued to evolve since the study took place.

when providers match participants to the model based on their existing work readiness.

Providers were all required to offer professional development sessions, but they varied in focus, content, and structure across agencies. Some sites integrated professional development throughout the work days, while others used Friday as a mandatory professional development day. Topics varied widely, ranging from developing business models to sexual health education.

In summers 2017 and 2018, a subset of WorkReady applicants entered a randomized lottery to allocate the limited number of slots. To minimize disruption to the city-wide program, only a small fraction of slots were assigned by lottery across both summers — about 12 percent in 2017 and 5 percent in 2018. A handful of providers were exempt from the lottery by preference or for logistical reasons, but the lotteried slots were generally representative of the program's work experiences as a whole. To facilitate cooperation among providers, PYN discouraged providers from serving control youth but did not prohibit it (if, for example, a control youth established a relationship with a provider after the lottery).

The experimental design differed depending on how youth applied to the program and the study year. In 2017, youth who applied directly to a local provider were pre-screened for eligibility, then randomly assigned within provider. Youth with no pre-existing provider relationships could submit online applications directly to PYN. This group is referred to as the "general pool" since they are not pre-screened. These applicants were blocked by geography and age, then individually randomly assigned within blocks to both providers and treatment or control groups. In 2017, 1,554 applicants across 39 providers entered provider lotteries, and 1,838 general pool applicants were randomly assigned across 20 providers.

In 2018, there was only one lottery consisting of 1,105 youth who applied to the general pool. PYN placed these treatment youth at 45 different providers according to provider needs. So in this cohort, there was no random variation in provider. In both study years, we randomized about 20 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. More detail on the experimental designs is in Appendix A.

3 Data

The data come from administrative databases capturing youth contact with various government agencies.[9] We use application and program participation data from the organizations in charge of administering each program: the Department of Family and Support Services in Chicago and the Philadelphia Youth Network in Philadelphia. In both cities, we use administrative school and police records to measure education and criminal justice outcomes. The data linkage used probabilistic matching, allowing for fuzzy matching on name, date of birth, and gender to attach study individuals to their police and school records in each city (see Appendix B for details).

Coverage for the Chicago data sources includes records from the early 2000s on. Coverage for Philadelphia arrest records begins in January 2012, so baseline arrest measures capture fewer years of data for the WorkReady sample (4.5 pre-randomization years for the 2017 cohort and 5.5 for the 2018 cohort). Both cover only arrests made by each city's police department. In the police data, we categorize each arrest as violent (a crime against a person), property (all theft, burglary, or larceny), drug (sale or possession), or other (everything else including vandalism, trespassing, illegal use of a weapon, warrant arrests, and other minor offenses) based on the offense's description. Youth who have never been arrested do not appear in the data, so we assign zero arrests for unmatched youth.[10]

[9] We also performed in-depth interviews with 18 WorkReady treatment and control youth in summer 2017. These interviews inform our description of the program, but were mostly used to help PYN understand variation in how youth experienced the program and set priorities for program adjustments (e.g., how youth are matched to job types, how to standardize the quality of participant-supervisor relationships, and how to increase opportunities for youth facing additional barriers to work).

[10] As is true in all studies using administrative police records, arrests are an imperfect measure of true offending behavior. They capture both police and individual choices. This generates both a downward bias in the measurement of crime, since many offenses do not result in arrest, as well as an upward bias, since not all arrests are for something an individual actually did. There is plenty of evidence that these biases do not affect all types of youth the same way, and likely vary systematically by race, neighborhood, and other characteristics of both the individuals and the arresting officers (Ridgeway and MacDonald, 2009; Goncalves and Mello, forthcoming; Hinton et al., 2018). One key benefit of the randomized design is that the study does not need to assume that arrests are a perfect measure of underlying criminal behavior; it is clear they are not. But the biases in the data-generating process affect both the treatment and control groups equally, and treatment effects measure the difference between the groups. Mismeasurement in the dependent variable might attenuate estimated treatment effects, but it does not bias them. Rather, the key assumption is that the treatment does not affect the probability of being arrested conditional on committing (or not committing) a crime. This is not entirely trival, since treatment could teach youth to interact with police more constructively, thus avoiding arrests that would otherwise have taken place. But even if that is driving some of the estimated treatment effects, the fact that criminal justice system involvement is so damaging

Though we have historical records from Chicago Public Schools back to the early 2000s, our data agreement with the School District of Philadelphia for this research only included records beginning in 2016-17. So we define baseline measures from the pre-randomization school year only, which we observe for everyone in both studies. Importantly, especially in the outcome data, missing education data is prevalent and substantively important. The absence of information is not random; missingness could indicate that students graduated, dropped out, transferred, or are attending a non-public school. Charter school enrollment generates considerable missing data in our context. In Chicago, charters report enrollment and attendance but not grades or disciplinary issues in the administrative records; in Philadelphia, charters only report enrollment consistently. This means grades, suspensions, and in Philadelphia, attendance data are missing for charter school students. This group makes up about 4 percent of the Chicago sample and a little over a third of the Philadelphia sample in the baseline years.

Because of the prevalence of missing data and the possibility that treatment could affect missingness by changing the probability of dropout, graduation, or transfer, we focus our outcome analysis on a single education measure that is non-missing for everyone who has any kind of education data available: school persistence, or an indicator for whether a youth has either graduated or remains enrolled in a public school as of a given school year. Those who have transferred, dropped out, or switched to non-district schools receive a 0 for this measure. We drop pre-program graduates from this analysis, since they cannot have changes in persistence by construction, as well as the 353 Philadelphia youth and 25 Chicago youth who did not match to school district records at all (presumably because they attended private schools or were not in school for the entirety of the study period). Appendix B and C.4 discuss and report results for other education measures with more missing data—absences, grades, and misconduct—using both non-missing data only as well as a variety of imputation methods. But the amount of missing data here, especially in Philadelphia where charters are so widespread, makes our interpretation of these results more cautious.

In Philadelphia only, the City's integrated data system, known as CARES, provides

to future individual and family outcomes (e.g., Aizer and Doyle, 2015; Mueller-Smith, 2015; Dobbie et al., 2018; Charles and Luoh, 2010; Holzer et al., 2006) means there still a large social benefit to reducing arrests, even if some of the change is not driven by changes in underlying crime.

information on a range of other outcomes that WorkReady might affect. We define indicators for other criminal justice interactions, including whether someone has been incarcerated in either juvenile detention or juvenile prison; we do not observe adult incarceration, so this variable is 0 for incidents occuring over the age of majority (18) or those charged as an adult. We also measure whether someone received any court-ordered services within the Department of Human Services (DHS), which is a less severe punishment than incarceration but can be offered to incarcerated youth as well. This variable understates the amount of total court-ordered services, since some services are provided outside of DHS; however, there is no reason to believe that treatment should affect which provider a youth is assigned.

We can also measure other social outcomes that are substantively important but less often available in individual-level administrative data. Using service records from the Child and Youth Division, we create an indicator for whether a study member's family received any services from the city's child protection agency. This includes any service in response to a substantiated call, which is one that the Division determines merits further involvement. It counts residential placements like foster care or kinship care, as well as non-placement services such as safety checks and other in-home services.

As with any service data, a change in services has multiple interpretations. It could represent a change in the number of incidents requiring child protective services; it could also reflect more or fewer incidents being called in or reported. Given that WorkReady involves more adult interaction with youth, it seems likely that conditional on a situation requiring intervention, treatment would increase the probability the incident is reported to the Child and Youth Division. If reporting increases for treatment youth, estimated program impacts on child endangerment would be understated.

To measure behavioral health, we have records of all substance abuse and mental health services that are either covered under Medicaid or funded in part by Philadelphia County.[11] We generate indicators for receiving any of these behavioral health services, as well as separately for mental health and substance abuse treatment. As with child protection, program effects on this measure could indicate changes in the underlying issue or changes in will-

[11]Due to HIPAA restrictions, the behavioral and mental health data that initially comes from medical records was provided to us separately from the rest of the data. The HIPAA-covered data is attached to fewer covariates to prevent re-identification, and it can not be linked to the rest of the study data.

ingness to seek out treatment. Since the most likely treatment effect of interacting with additional caring adults would be to encourage the identification of personal issues and willingness to seek help, we expect that this measure may also understate any treatment-driven declines in the true underlying health conditions.

Lastly, we create indicators for whether a youth ever used a city homeless shelter and whether they had a child. Fertility is measured as being listed on a birth certificate in Pennsylvania vital statistics records, so there is more scope for missing data and treatment effects on reporting for fathers than mothers. All other measures in the paper are separated by year since random assignment. But to protect confidentiality, the City only provided information on whether births were pre- or post-random assignment. As a result, we do not separate the parenthood outcome by year; it is measured across all observed post-randomization years. Together, the social services data provides a much more complete picture of youth and family welfare than has been available in prior summer jobs studies. The data allow us to capture, for the first time in administrative data using experimental methods, the programs' influence on family income, stress, and stability; self-efficacy and mental health; and risky behaviors like substance use or unprotected sex.

4 Analytical Methods

We estimate the intent-to-treat effect with the following ordinary least squares regression:

$$y_{ibt} = \beta_0 + \beta_1 T_{ib} + B_2 X_{ibt-1} + \gamma_b + \varepsilon_{ibt}$$

where Y_{ibt} is the outcome of interest for individual i in randomization block b in period t. T_{ib} is an indicator for individual i being randomly assigned to be offered a program slot. X_{ibt-1} is a set of individual i's pre-randomization characteristics, and γ_b is a vector of randomization block fixed effects (see Appendix C for list of baseline covariates). For any missing baseline covariates we impute 0s and include an indicator for missingness. We show results with no baseline covariates other than the block fixed effects (and for OSC+, duplicate application indicators) required for identification as a robustness check in Appendix C. To ease interpretation, our main analysis uses ordinary least squares. Since outcomes are either

indicators or counts, Appendix C.2 reports average marginal effects from logistic or Poisson regression (with robust standard errors to relax the assumption that the mean and variance are equal); substantive conclusions are unchanged.

The ITT estimates the effect of receiving an offer to participate in a summer jobs program. Since not all treatment youth and some control youth participated in the program, the ITT will understate the effect of actually receiving program services. To estimate the effect of actually participating on compliers, we use random assignment as an instrument for ever participating, defined as having more than 0 hours recorded in program records. Given the control crossover, this estimator is formally a local average treatment effect (LATE). To assess the magnitude of these effects, we report estimates of control complier means (CCMs) as a baseline.[12]

In addition to reporting heteroskedasticity-robust standard errors, clustered on person in WorkReady where 132 applicants appear in both cohorts, we also conduct randomization inference to test the sharp null of no program effects for anyone in the sample (Athey and Imbens, 2017; Fisher, 1935). Appendix C.3 reports these adjustments, as well as inference adjusting for multiple testing by controlling either the family-wise error rate or the false discovery rate (Anderson, 2008; Westfall and Young, 1993; Benjamini and Hochberg, 1995). We adjust within families of outcome types: the different overall measures of criminal justice involvement (incarceration, juvenile justice services, and total arrests, available in Philadelphia only); the type of arrests (violent, property, drug, and other in both studies); family outcomes (child protection services, shelter use, and fertility); and behavioral health (substance abuse and mental health services). We do not adjust testing for education since there is only one main outcome (persistence) in that family.

To test whether there is significant treatment variation across providers, we use the subset of the data with random variation in provider assignment (all of Chicago and the 2017 general pool lottery in Philadelphia). We include baseline covariates, block fixed effects, and provider-specific random effects on the treatment indicator, then use a likelihood ratio test to assess whether the model allowing variation by provider statistically differs from the fixed

[12]Given the low baseline means of many outcomes, estimates of CCMs for indicator or count variables are sometimes negative due to the sampling error in the LATE. We round these cases to 0.

treatment effect model.[13] Since provider assignment is random within block, the regression does not require the inclusion of provider fixed effects; however, their inclusion does not change the results. Each block with random provider variation contains 2-4 providers.

To test for heterogeneity by risk, we implement both the leave-one-out and repeated-split-sample procedures in Abadie, Chingos and West (2018). The details of this approach are in Section 7.2.

5 Descriptive Statistics and Compliance

5.1 Sample Composition as Programs Scale

Table 1 shows baseline characteristics for both the WorkReady and OSC+ study populations prior to random assignment, as well as tests of treatment-control balance for each covariate. No more of the differences are significant than would be expected by chance,[14] and as shown in the last two rows, the tests of joint significance in both studies confirm that treatment and control groups are balanced. It is worth noting that one of the chance imbalances in the WorkReady study is on the primary pre-specified outcome, with the treatment group having fewer pre-program violent-crime arrests ($p = 0.01$). Although imbalance on this outcome is unfortunate, the difference is controlled for by including baseline covariates in all outcome regressions. Additionally, as discussed below, the overall level of violent-crime arrests (and in fact, all arrests) ended up being lower than expected at the outset. So the results separated by crime type are less informative than expected at the time of pre-specification.

One of the challenges to scaling programs is whether the type of youth served changes as the program grows. In the case of OSC+, we can observe those changes directly by comparing the characteristics of this OSC+ applicant population to the initial 2012 cohort.

[13] In theory, including provider-by-treatment interactions and testing whether they are all equal is another option for this test. But estimating the separate provider fixed effects adds uncertainty from the estimation of the fixed effects, reducing the power to distinguish effects across providers. Since we have random variation in provider assignment, the assumptions for random effects are met by construction; provider assignment is not correlated with any other covariate, conditional on block. So we focus on the random effect approach to aid power.

[14] For WorkReady, 2 of 26 tests have $p < 0.05$, about what would be expected if all covariates were independent (which they are not, since some are sums of other covariates shown). The joint tests at the bottom of the table excludes variables that are linear combinations of the others.

Eligibility criteria were largely similar across those study years,[15] but the number of available program slots roughly tripled.

In fact, the baseline characteristics of the OSC+ population look fairly similar to the characteristics of those in the initial study. As shown in Table 1, applicants were just over 17 years old on average and about 40 percent male. Twenty-two percent had been arrested at baseline. Because of the school-based eligibility rules, basically the whole population was in school prior to the program. On average, applicants were between 10th and 11th grade, with about a quarter graduating at the beginning of the program summer. They had an average GPA of 2.4 and missed 24 days of school. For comparison, the 2012 applicants were a little younger (16.3 on average, because 14- and 15-year-olds were eligible then), but also about 40 percent male, 20 percent with an arrest record, missing about 33 days of school, with GPAs averaging 2.4 (Davis and Heller, 2020; Heller, 2014).

The similarity of observable characteristics across study cohorts demonstrates that even the tripling of the applicant pool between the 2012 and 2015 studies did not dramatically shift the make-up of the applicant pool. Maintaining similar youth populations as programs grow may not be feasible in every setting. But at a minimum, in a large city like Chicago, providers were able to identify and recruit a broader group of youth without much change in school engagement or criminal involvement.[16] The one major difference between samples is that the current study participants were about 23 percent Hispanic, but only 3 percent of the initial study. Given the residential segregation in Chicago, this change is likely due to the expansion of the program into more Hispanic areas.

Table 1 also shows baseline characteristics for the WorkReady sample. About 80 percent self-identified as African-American and 12 percent Hispanic. The Philadelphia program

[15]The 2013 cohort involved explicitly different eligibility criteria, recruiting only males, some from the criminal justice system, to test for effect heterogeneity. By contrast, the 2012 recruitment worked like the current study, recruiting anyone attending an eligible school. But the initial study had 700 program slots spread across 13 schools, while this study had 2,000 program slots spread across 49 schools.

[16]As a rough benchmark for how big the target population of those who might benefit from an SYEP that reduces arrests is relative to the 5,405 applicants, around 13,000 people under age 17 were arrested in Cook County, where Chicago is located, in 2015 (Gleicher, 2017). The ACS reports a little over 150,000 people between 15 and 17 living in Chicago. So while there is likely to be continued scope for program expansion without major changes in the participant population, a universal program for everyone under 17 would likely dramatically change the criminal justice involvement of the participants. This calculation emphasizes the importance of thinking carefully about the target population when making decisions about how large programs should be.

included somewhat more white, Asian, and other race categories than in OSC+. But as in many cities, both SYEP populations disproportionately include non-white youth relative to city populations.[17]

Although the study sample is not representative of the WorkReady program as a whole (applicants without pre-existing relationships with providers are over-represented, see Appendix A), it does reflect some differences in the scale of the two programs and whom they target. WorkReady applicants were almost two years younger than in OSC+, with an average age just under 16. The WorkReady study sample was considerably less involved in the criminal justice system than in Chicago, with only about 4 percent having an arrest record prior to randomization. This could in part reflect the fact that Philadelphia arrest records only begin in 2012 and so have a shorter coverage window. But it likely also reflects that OSC+ targets youth at high risk of violence involvement by design, with many other programs in Chicago serving less criminally-involved youth. WorkReady, on the other hand, is the main city-wide SYEP, so was not targeting applicants more likely to have prior criminal justice contact. A similar pattern appears in the school data, with WorkReady applicants missing 18 days of school relative to OSC+'s 24 days. Similarly, 15 percent of those in the WorkReady school data had been suspended in the year before random assignment relative to 20 percent of OSC+, although GPAs are more similar across studies.

Additional data availability in Philadelphia also provides a broader description of the level of disadvantage facing the WorkReady sample. About 14 percent of the study were in families that at some point received child protective services, some when they were very young. About 4 percent had stayed in a homeless shelter; 1.6 percent were parents themselves; and about 26 percent of the sample had received behavioral health services, most consisting of mental health care. These rates are about 50 percent higher than the population of youth in the City's service database that did not apply to WorkReady.

[17]American Community Survey data shows that in 2017 Philadelphia was about 35 percent white, 41 percent Black, and 14 percent Hispanic; Chicago in 2015 was 32 percent white, 31 percent Black, and 29 percent Hispanic. So Black youth are highly over-represented and white youth are under-represented in applicant pools for both cities, with Hispanic youth applying slightly less than proportional to their population.

5.2 Compliance

Both studies faced compliance challenges. In Chicago, this was due to an error in a new online system that the City implemented to transmit lists of lottery winners to program providers. During the initial recruitment period, the city's contracted programmer unintentionally allowed unrestricted access to all applicants, listing the control youth after the treatment youth rather than withholding the waitlist from view. Because of this error, agencies could initially click through to view all of their control group applicants, though not all of them did so prior to the research team catching the error. Additionally, as had been the case in the past, not all treatment youth could be reached by their assigned agency or were still interested in participating. The resulting first stage for OSC+ is 0.26 (F-stat = 454), with 46.5 of the treatment group and 20.4 of the control group participating.

In Philadelphia, non-compliance was the result of a strategic choice by the non-profit agency operating the program, PYN. To increase provider participation in the new lottery system and the corresponding broader outreach to new youth populations, PYN encouraged but did not force providers to adhere to random assignment. In the first study year, 44.5 percent of treatment youth and 18.5 percent of control youth worked at least one day, for a first stage of 0.26. In the second year, PYN worked hard to reduce some of the barriers providers faced in serving youth with whom they had no pre-existing relationships. They held recruitment sessions where PYN staff helped youth fill out the required paperwork for the program, rather than relying on providers to do so on their own. And they extended personal invitations to treatment youth to encourage their attendance at these sessions. Combined with the fact that the 2018 study focused solely on a general pool lottery (so providers did not know the identity of control applicants), this strategy successfully increased take-up to 67 percent among the treatment group and reduced it to 9 percent among controls, for a first stage of 0.58. Pooling the two cohorts, the first stage is 0.34 (F-stat = 604).

6 Results

This section begins by discussing impact estimates for outcomes that are available across both cities: criminal justice involvement and education. It then turns to results on other

family and health outcomes, which are only available in Philadelphia, followed by tests of how variation in program structure and delivery matter. Both sets of results confirm the promise of SYEPs to have similar effects across multiple scales and settings.

6.1 Crime and Education

Table 2 shows the estimated ITT and LATE for criminal justice involvement and school persistence in the first year after randomization. In both cities, the programs generate proportionally large decreases in contact with the criminal justice system. In Philadelphia, being offered the program results in 1 fewer arrest per 100 youth, a statistically significant 36 percent decline. Due to a first stage far less than 1, the effect on compliers is much larger: 3 fewer arrests per 100 participants, a 65 percent decline relative to the CCM. The decline also translates to a decrease in juvenile incarceration. Although the latter result is marginally significant ($p = 0.09$), it is also proportionally huge: Incarceration drops by 1.5 percentage points, or almost 80 percent among WorkReady participants.

The point estimate on total arrests is larger in levels but proportionally similar in Chicago. OSC+ participants have almost 9 fewer arrests per 100 participants than control compliers (a 52 percent decline), although the result is not quite statistically significant ($p = 0.125$). Across both cities, there are statistically significant and substantively large declines in drug and other arrests, ranging from 20 to 100 percent drops relative to baseline means.[18] The point estimates on violent and property crime are similar in magnitude, negative, and proportionally large in Philadelphia, but not statistically different from zero. Where data are available for longer-term follow up (see Appendix D), WorkReady significantly decreases property crime by about two-thirds in the second year post-random assignment (0.6 fewer arrests per 100 youth offered the program relative to a control mean of 0.9). There is a similarly large and statistically significant decline in property and other crime arrests in year 3 for OSC+ (ITT of -0.6 per 100 youth relative to a control mean of 1.4 for property crimes, and -1.3 per 100 relative to a control mean of 6.2 for other crimes), generating a drop in total year 3 arrests of 2.2 per 100, a 20 percent decline. The pattern of effects previews the risk-responsiveness relationship we investigate below: Outcomes with larger control means

[18]The drug arrest result in Philadelphia is just barely above the standard significance threshold, $p = 0.104$, which we treat as marginally significant.

also show larger point estimates.

A number of the main results cross traditional significant thresholds after adjustments for multiple testing (see Appendix C.3). Yet this seems more likely due to the power limitations that come with non-compliance than a serious risk of Type I errors. The probability that both cities' results are Type I errors is considerably lower than the probability either one is in isolation, so the built-in replication across sites should increase confidence in the results. And the fact that criminal justice involvement has fallen in all previous SYEP studies also strengthens confidence in the result, despite the individual p-values rising slightly above the 0.1 cutoff after multiple testing adjustments.

There are no significant changes in school persistence, defined as either remaining enrolled in school or having graduated. Point estimates are small, ruling out more than a 2 percent increase for those offered the program in either city. Because of the amount of missing data on other education outcomes, discussed in section 3, it is more difficult to estimate effects on other education outcomes with confidence (see Appendix C.4 for results and discussion).[19]

These results are a confirmation that SYEPs consistently reduce criminal justice involvement, even when they are scaled up or implemented in different contexts. WorkReady as implemented at scale across all of Philadelphia reduces arrests and keeps young people out of detention and prison. Given that prior studies in Chicago and Boston focused on small, selected subsets of program providers, and that prior studies in New York examine a version of the program (and of the criminal justice system) that is now 15 years old, this is an important replication. OSC+ also continues to reduce arrests, despite having a different supplementary curriculum, different mix of job types, lower wages relative to the labor market, and a new set of program providers relative to previous studies. In other words, variation in program details and delivery seems to have little effect on impacts. We return to this point below in testing experimental variation in program structure and delivery.

It is worth noting that the type of crimes that respond to programming here differs from prior studies of OSC+. In prior work (Heller, 2014; Davis and Heller, 2020), OSC+ crime

[19]Appendix Tables A6 through A9 show no clear improvements in other education outcomes, regardless of how missing data are imputed. If anything, the point estimates tend to be negative. Out-of-school suspensions may have increased in the OSC+ treatment group, and year 2 school persistence decreased, but not in WorkReady. Prior work in Chicago has not used information on misconduct because of data quality concerns; see Appendix C.4 for discussion.

declines were all driven by decreases in violent-crime arrests (which is why violence was the primary pre-specified outcome in our pre-analysis plan). While it is possible the shift has to do with the substantive changes in programming, it is also the case that the overall level of violent-crime arrests in the control group is much smaller in the current studies (1.1 and 3.7 per 100 control youth here, relative to 7.4 per 100 in the 2012 study and 10.8 per 100 in the 2013 study).[20] Indeed, the 30-40 percent declines in violence found in prior studies are within the confidence intervals of both the WorkReady and OSC+ studies. So although there is not enough precision to confirm a similarly-sized decline in violence in these studies, we can not rule it out. Regardless, the overall pattern of effects on arrests shows a consistent reduction in criminal justice exposure.

6.2 Other Family and Health Outcomes

SYEPs provide income, which could reduce family or individual stress. The extra $1-2,000 could improve living conditions at home; in surveys of Chicago participants, almost 80 percent of net wages went to either local businesses or participants' families (MHA Labs, 2015). SYEPs also aim to develop personal skills and self-efficacy, which could directly shape mental and behavioral health, as well as risky behaviors. Yet prior work on the behavioral effects of these programs has largely been limited to crime and education outcomes.

Table 3 shows WorkReady's impact on other family and health outcomes in the first year after randomization. There are some potentially promising results. Among those offered the program, there is a marginally significant ($p = 0.09$) 0.6 percentage point, or 33 percent, decline in the receipt of child protective services, which include in-home safety checks, other family services, and removal of children to residential placements. The LATE is a 1.6 percentage point decline among compliers, which is basically a 100 percent decline relative to the CCM. There is reason for caution about the strength of these results; they are proportionally huge changes, but from a very low baseline, and they do not survive adjustments

[20] Given the similarity in number of arrests prior to random assignment, the drop in violent crime involvement among controls does not seem to be due to fundamentally different risk levels. It likely is due partly to the large secular drops in violent-crime rates citywide over time, as well as changes in policing that decreased arrest rates among youth (see, e.g., `https://home.chicagopolice.org/statistics-data/statistical-reports/annual-reports/`). So it is possible that the lack of a decline in violence is because the lower occurrence and recording of violent events makes behavioral changes harder to detect in the data.

for multiple testing within this family of outcomes (see Appendix C.3). Nonetheless, there is more precision in some subgroups (including for African-American youth, see Appendix E) and when using a probit, which may handle the low base rate somewhat more effectively than the linear probability model (average marginal ITT effect of 1 percentage point, a 45 percent decline, p = 0.028, see Appendix C.2). Because intervention due to a substantiated call about concerns over child safety is an extreme and costly outcome, even suggestive evidence that WorkReady reduces these services merits attention.

Similarly, although the point estimate on the receipt of behavioral health services is not statistically significant (p = 0.207), it is proportionally quite large. Youth who participated in WorkReady were 3.3 percentage points, or 31 percent less likely to receive these services. As shown in the bottom two rows, most of these services are mental health related, with a little under 1 percent of them addressing substance abuse issues. This is certainly not strong enough evidence to conclude that SYEPs reduce behavioral health problems with confidence. But as discussed in Section 3, program effects may be understated; to the extent that WorkReady increases the probability of detecting and reporting family or health issues, service receipt might increase conditional on the underlying issue for treatment youth, attenuating estimated program effects. Some of the subgroup results in the appendix also point toward the possibility that SYEPs may matter for family and health outcomes, especially for subgroups like boys and African-American youth who are at elevated risk of the outcomes. The direction and magnitude of these effects for outcomes that are central to youth well-being should be a priority for future study.

6.3 Variation across program structure and delivery

Because variation in program model and implementation are two key hypotheses for why other interventions do not always replicate across contexts, the research design included experimental variation in both features. In Chicago, we randomly varied the structure of the program across two treatment arms — a job only versus a job, mentor, and civics curriculum. Table 4 shows no significant differences by arm, although the pattern is suggestive: Only the job + mentor group has crime declines that can be differentiated from 0.[21] The point

[21] Because we do not have separate measures for who actually worked without a mentor and who received real mentorship, we can not instrument separately for the two types of activities. As a result, the table

estimates on arrests for the mentored group are generally about twice as large as in the job-only group, providing some hint that the additional program elements do increase the program's impact. And in year 2, school persistence declines among the job-only group (see Appendix C.4). This is suggestive evidence that the extra supports may help prevent youth being pulled into the labor market before finishing school. But there is not enough statistical power to differentiate the two groups.

Random assignment to provider facilitates a different kind of test of how much program structure and delivery matter. As described above, in both OSC+ and the 2017 general pool lottery, the experimental design means provider is uncorrelated with youth characteristics and treatment probability, conditional on randomizaton block. We test for cross-provider heterogeneity by including random treatment coefficients (i.e., random slopes) for each provider, then testing whether the random effects explain significantly more variation than a single treatment indicator. There is no detectable variation in treatment effects across providers. Across both studies, for only 1 of the 17 main year 1 outcomes can we reject the null that the random effects have no additional explanatory power relative to the single fixed treatment effect.[22] It is possible that the block fixed effects, which capture geographic variation in both cities and age variation in Philadelphia, absorb some of the differences across providers, or that the 19-20 providers in each city's test are not enough to generate detectable variation. Nonetheless, the lack of significant variation across providers suggests that the crucial part of the program may be in the overall approach, not the details of job placement, program staff, or organizational structure. This is somewhat rare in human capital interventions, where implementation details often seem to matter, and it likely contributes to the program's replicability and scalability.

7 Variation by individual risk level

Heterogeneity by risk, defined as the level of some counterfactual outcome Y_0, is of interest in part because the distribution of gains by baseline risk level informs questions of equity

focuses just on the ITT effects by arm. Take-up was slighly higher in the mentor arm, with a first stage of 0.29 in the mentor group and 0.24 in the job-only group. But the difference is not statistically significant (p = 0.25).

[22]The significant variation is for drug-crime arrests in Chicago. One out of 17 is about what we would expect by chance.

and social justice (see e.g., Heckman et al., 1997). If those whose behavior changes the most in response to intervention are not those at the lowest part of an outcome distribution, there may be tradeoffs between equity and efficiency in deciding whom to serve. More broadly, the shape of the relationship between Y_0 and treatment effects may also reflect something deeper about the nature of heterogeneity that has important implications for optimal targeting. Debates about the merits of prevention versus remediation are, in part, a question of whether some level of Y_0 is so high as to prevent responsiveness to the treatment.

A typical approach to heterogeneity is to search for how variation in X, rather than variation in Y_0, affects β. This can be effective if there are a small number of observables that drive large differences in treatment heterogeneity, but it has limitations. Most studies are powered to detect main effects but not subgroup effects; searches over multiple interactions further reduces power by generating the need for multiple testing adjustments; and more flexible machine learning approaches require additional sample splitting to get inference right. So subgroup analyses are often quite under-powered. Plus, within any single study, it is difficult to tell whether a particular characteristic drives bigger treatment effects because something about the group causes a differential treatment reponse, or whether that X just happens to be correlated with something else about the setting that matters, such that targeting the group in a different setting would not be as effective. By focusing instead on the direct relationship between Y_0 and β across settings, we aim to draw some broader lessons about patterns of treatment heterogeneity.

To elucidate how the relationship between Y_0 and β matters for decisions about investing in a new or expanding program, this section begins with a conceptual framework for thinking about the risk-responsiveness relationship and targeting decisions. It then uses impact estimates across multiple outcomes and studies to estimate the shape of that relationship and discusses what we learn from the results.

7.1 Framework relating heterogeneity, risk level, and targeting

Consider the set of outcomes, a vector $\mathbf{Y}$, that SYEPs affect. Across existing studies, these are almost always counts or indicators for crime or harmful health and welfare outcomes. So define $\mathbf{Y}$ such that all elements $Y_i \geq 0$, and decreases in each Y are socially beneficial.

Different types of people, indexed by θ, may respond differently to treatment. So in a potential outcomes framework, each row of the treatment effect vector, $\beta_{\mathrm{Y},\theta} = E[Y_1(\theta)] - E[Y_0(\theta)]$, may vary by group and by outcome.

Each occurrence of each Y has an associated social cost, C_Y. All else equal, policymakers considering how to generate the most social benefits with SYEPs (or any program) would like to target the θ groups who make $\beta'_{\theta,Y}C_Y$ as negative as possible.[23] Because each occurrence of Y is socially costly, a social planner would want to maximize the number of events prevented, weighted by cost. In other words, the absolute magnitude of the treatment-driven decreases in counts matters more than the size of the proportional changes; 3 fewer events from a baseline of 12 is more socially beneficial than 1 fewer event from a baseline of 1 (i.e., $\mid -3C_Y \mid > \mid -C_Y \mid$), even though the former is a 25 percent change and the latter is a 100 percent decline. A detailed dive into the social costs of the different behaviors — crime by type, child abuse and neglect, mental health and substance abuse, etc. — involves specifying a social welfare function, requiring normative judgments beyond the scope of this paper. So here we focus on what the data can tell us about the type of youth with the most negative $\beta_{\theta,Y}$, which is a key input, if not a final answer, to optimal targeting decisions.

To make the implications of the risk-responsiveness relationship for targeting concrete, consider treatment heterogeneity across variation in some counterfactual outcome, Y_0. Figure 1 shows three stylized examples of how responsiveness to treatment could vary by risk of this outcome, plotting theoretical variation in β_Y across a population with different levels of risk of the outcome in the absence of the program, $Y_0(\theta)$. Each panel represents a different structure of treatment heterogeneity. Panel A shows a case where treatment shifts the outcome down by some constant amount α for everyone, regardless of their risk level. Across most of the distribution, β_θ is a constant, $-\alpha$, for any choice of θ and the corresponding Y_0. The exception is for very low risk individuals whose $Y_0 < \alpha$. Since Y_1 can not be negative, there is a floor effect, with β getting smaller when the population served is at almost no risk of the negative outcome. Here, policymakers would generate equivalent social gains

[23] In practice, all else might not be equal. For example, the cost of serving different types of individuals may vary, such that policymakers would need to balance bigger benefits with higher costs. Or the social costs of a behavior could also vary by θ, as would be the case if policymakers cared about the distributional impacts of a program. We return to this point in the discussion below.

regardless of which population they served, as long as they chose θ such that $Y_0(\theta) > \alpha$. Panel B, by contrast, shows a case where the treatment effect is a proportional shift in the outcome, $\beta_{Y,\theta} = -\alpha Y_0(\theta), 0 < \alpha < 1$. Here, policymakers should want to serve individuals as far to the right of the graph as possible; bigger $Y_0 s$ correspond to bigger social gains.

Panel C shows a more complicated case, motivated by the idea that behavior may only change on the margin. Suppose, for example, that those very deeply involved in crime are committed enough to their behavior that a summer intervention would have little effect. And suppose that those barely involved in crime commit offenses rarely enough that an intervention is unlikely to matter much. In this case, serving types of people with high *levels* of the outcome in the absence of the program is not the same as serving people with big *changes* in outcome due to an intervention. It is only participants in the middle whose behavior might be shifted, those who are close enough to the margin of crime for a time-limited intervention to change their decisionmaking.[24] Here, policymakers should try to identify those for whom $Y_0(\theta)$ is in the responsive region, ideally close to the peak of the β_θ function. Of course, these are stylized examples. There are many possible forms the relationship between Y_0 and β could take. The point is that the shape of the relationship matters, both for targeting choices and for understanding the nature of the behavioral response more broadly.

7.2 Estimating effect heterogeneity by baseline risk level

Estimating this relationship empirically is challenging, because Y_0 is not observed for the treatment group. We take two different approaches to understanding the relationship between Y_0 and β_θ.[25] First, we perform an endogenous stratification exercise. Abadie et al. (2018) show how to use the relationship between the Xs and Y_0 in the control group to

[24] One interpretation of the effects of active labor market programs more broadly is that they follow this kind of pattern. Many short-term programs that target those with the highest barriers to employment, like the long-term unemployed or those returning from prison, have historically had mixed to no effects (Berk et al., 1980; MDRC, 1980; Cave et al., 1993; D. Bloom, 2010; Heinrich et al., 2013; Doleac et al., 2020; Card et al., 2017). Some, like the JTPA, even have adverse crime effects on those already at elevated risk of crime (H. S. Bloom et al., 1997). Many of the more recent programs that do have large, positive employment effects perform purposeful screening upfront, potentially finding those close to the margin of success but still at risk of a bad outcome as a way to ensure programs effectively help participants cross the relevant margin (Fein and Hamadyk, 2018; Schaberg and Greenberg, 2020; Roder and Elliott, 2020). Within-study heterogeneity is also consistent with this idea, as in the classic Friedlander (1988) finding that the biggest responses occurred among those who were in a middle-risk tier, neither too well off nor too disadvantaged at baseline.

[25] Note that this analysis was not pre-specified, so should be considered exploratory.

predict Y_0 for the treatment group, then estimate heterogeneous effects across the predicted risk groups. The procedure involves regressing the outcome on all the baseline covariates using just the control group, predicting Y_0 for the treatment group using those regression coefficients, separating the observations into groups by their level of $\hat{Y_0}$, then estimating separate treatment effects for each group. To avoid the finite sample bias that comes from fitting a prediction regression within sample, the authors suggest using both leave-one-out regression and repeated split samples.

We note upfront that predicting a single Y_0 and estimating treatment heterogeneity by $\hat{Y_0}$ has practical limitations. The adjustments required to avoid bias in finite samples reduce power, even more than a typical subgroup test. The approach typically estimates average treatment effects for two or three parts of the $\hat{Y_0}$ distribution. It is difficult to extrapolate the full shape of the risk-response distribution from two or three points. And given how many different outcomes SYEPs seem to affect, we may re-introduce multiple testing concerns by repeating this exercise as many times as available outcomes.

To address some of these issues, we perform the heterogeneity test with an index that combines information about the underlying risk of socially costly behavior across measures. Although this risks masking heterogeneity that varies by outcome, it has the benefit of increasing power by combining information on outcomes that tend to move in the same direction, and reducing the number of hypothesis tests (see, e.g., Kling et al., 2007). To do this, we standardize each outcome with a statistically significant main effect and generate an unweighted average of these outcomes by city.[26] In Chicago this includes drug and other arrests; in Philadelphia it includes incarceration, total arrests, and child protective services.[27] Each variable is standardized on the control group, averaged together, then re-standardized so that the standard deviation of the index is 1.

Table 5 reports the overall ITT effect on the index, a 0.065 standard deviation decline

[26]In our context, a joint analysis across studies is logistically impossible. The data are not all held in the same place; Chicago and Philadelphia data are on separate institution's servers due to data agreement limitations. Even within Philadelphia, the behavioral health data are completely separate, with very limited Xs available, to comply with HIPAA regulations. So we can not include behavioral health data in this exercise. We limit the index to statistically significant main effects to avoid diluting the index with additional noise. Appendix F shows that we get a generally similar pattern of results when using an index of all the available measures of socially costly outcomes, clearer in Philadelphia than in Chicago, but with less precision.

[27]We exclude drug and other arrests since they are already part of the total arrest outcome.

in Philadelphia and 0.051 decline in Chicago, as well as the results of the endogenous stratification exercise.[28] Consistent with the pattern in the main crime results, treatment point estimates are considerably more negative for the groups with higher predicted baseline index values. In the Philadelphia repeated split sample estimation, the effect for the high-risk group is more than 24 times as large as for the low-risk group; for the leave-one-out estimation, the change is even starker as the low-risk coefficient flips sign. This is an ITT analysis, so the treatment effect differences capture both differences in responses and take-up rates. Interestingly, the first stage declines considerably as risk rises, from 0.41 in the low-risk group to 0.28 in the high risk group. So the actual differences in responsiveness conditional on take-up are even larger than the differences in ITT point estimates would suggest. A similar pattern occurs in Chicago, with point estimates 13 – 26 times larger for the highest risk group than the lowest. The decline in take-up across risk groups is more muted here, a difference likely linked to the cities' different recruiting strategies (see discussion in the companion project, Bhanot & Heller, in progress).

Yet the practical limitations of the endogenous stratification approach are clearly reflected here. Despite being quite large changes relative to the control means, the size of the standard errors makes it difficult to differentiate the groups from each other. And although it looks like treatment effects might be growing across groups, the pattern varies somewhat across the two estimation strategies, especially in Philadelphia. So it is not clear whether effects grow proportionally across groups or are just concentrated among the highest-risk group.

The lack of power to clearly understand the pattern of treatment heterogeneity is a common problem within any single study. And it leads to the second approach, which is to consider what we can learn from variation in risk level and heterogeneity patterns across groups in different studies. To do so requires stepping back from the individual-level variation in a given Y_0, and instead focusing on group variation across different $E(Y_0)s$. That is, in the spirit of meta-analysis, we compare treatment effect variation across groups that have varying average outcomes in the control group. SYEPs provide an unusual opportunity to do so, because there are now so many experimental treatment effects available across the

[28] We rely on the user-written Stata command estrat for this analysis, with a small adjustment to the code to ensure that results are exactly replicable within the same seed.

two separate experiments in this paper, plus published main and subgroup effects from two prior OSC+ cohorts, NYC, and Boston experiments. We use 62 different significant point estimates, across all socially-costly outcomes, study locations, and subgroups, to look at how treatment effects vary across existing, measurable variation in $E(Y_0)$, estimated by $\bar{Y}_0$.[29]

This approach also has limitations. By combining information across different outcome measures, this kind of synthesis makes it difficult to draw specific conclusions about mechanisms from the heterogeneity patterns. The numerical slope of the function relating $\bar{Y}_0$ and β_{Y} is not directly interpretable, since a one-unit increase in $\bar{Y}_0$ represents different kinds of increases in the prevalence or count of different socially costly outcomes.[30] And in some ways, it is a weak test of the relationship between the elements of $\beta_{\mathbf{Y}}$ and $\mathbf{Y_0}(\theta)$; if there is no clear relationship, it may just be because the various outcomes making up the $\mathbf{Y_0}(\theta)$ vector have different risk-responsiveness relationships, such that aggregating them masks patterns in the individual outcomes. On the other hand, if there is a clear pattern in how treatment effects vary when a set of outcomes is more or less common in a group, then using so many data points could be a productive way to gain insight into the shape of the relationship between the risk of socially costly outcomes in a population and the magnitude of the change an SYEP generates in that population.

Figure 2 plots each group's control mean on the x-axis against the corresponding LATE estimate on the y-axis, using estimates that are individually significant at the $p \leq 0.1$ level. Though the control mean is not the most relevant baseline comparison for the compliers who drive the LATE, most other SYEP studies do not report control complier means. So using the control mean rather than the control complier mean allows us to show the same set of relationships across studies, and it still captures the basic relationship between the risk level

[29] Appendix F lists which estimates are included in each panel. To ease interpretation, we focus only on the outcomes where negative effects are desirable. One could also do this exercise with the positive educational effects in Leos-Urbel (2014), Schwartz et al. (2015), and Modestino and Paulsen (2019), but we avoid that here in part because those results differ from the education effects in this paper, while the other results are more consistent across studies.

[30] Standardizing the outcomes could be a partial solution. But since so many outcomes are indicator variables, focusing on mean changes rather than standard deviation changes is more directly interpretable. And importantly, switching to standard deviation units would dramatically limit our ability to include estimates from other studies, since none of them report outcome standard deviations, either overall or by subgroup. Regardless, since all outcomes are either indicators or counts, the units are still roughly comparable: An increase of 0.01 for any of the outcomes reflects 1 extra occurence of a negative incident in a group of 100 youth offered the program.

of a group and its response to treatment (inclusive of take-up decisions). Appendix Figure A.1 shows a similar relationship using the ITTs from each study, with a little extra variation induced by the differences in take-up rates across subgroups.

Panel A of Figure 2 starts with the significant main effects in this paper across outcomes and cities. This focuses on the variation in $\bar{Y}_0$ that comes from the different city populations, as well as the prevalence of the different outcomes. The panel plots each LATE point estimate against the corresponding control mean.[31] The pattern is strikingly linear; larger control means are consistently associated wth proportionally larger SYEP-driven declines the outcome. This suggests that among the outcomes responsive to the program, SYEPs have a bigger effect for groups that are more likely to be at risk of those outcomes.

To further explore the robustness of this pattern, Panel B adds point estimates and control means from the full study populations in previously-published studies of SYEPs.[32] This adds more variation in the prevalence of each outcome across independent populations. Despite differences in programming, time periods, and local context, the relationship is quite consistent. Increases in control means are roughly linearly associated with more beneficial treatment effects. The same also appears to be true for the few adverse effects (the positive green diamonds are both increases in later property crime from the initial OSC+ study), with bigger control means linearly associated with positive effects as well.

Panel C adds significant effects by subgroups across studies, reflecting variation within each outcome and study driven by a single division on one observable characteristic at a time. Appendix E presents and discusses the substantive subgroup results for WorkReady and OSC+ 2015.[33] Here we focus on the overall pattern between subgroup differences in

[31]Both here and below, point estimates in the plot are not always independent. For example, some arrest categories are included in the total arrest category, and some subgroups compose part of the overall estimates from the same study. But we avoid plotting estimates that are purely linear combinations of each other (e.g., if we show effects for OSC+ 2012 and 2013 studies separately, we do not also show the pooled estimate).

[32]This includes crime, incarceration, and mortality effects in Davis and Heller (2020), Modestino (2019), and Gelber et al. (2016). The Boston paper reports only the ITT effects; we use the reported take-up rate to scale the ITT, backing out the LATE.

[33]We are cautious not to over-interpret any given subgroup estimate given the number of hypothesis tests in all these interaction effects. A few findings may merit further attention in future work powered to distinguish subgroup effects. 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, and a significant overall drop in combined behavioral health services. African-American youth show a large and significant drop in child protective services (ITT = 0.9 percentage points, a 43 percent decline). Males also show the only significant adverse effect, a 0.9 percentage point increase in parenthood, tripling

baseline rates and subgroup responsiveness.

Across subgroups that vary in their risk of these outcomes, the size of treatment effects still seems to scale proportionally with the size of the control means. There is perhaps a bit of flattening in the middle of the graph, but overall, the declines in costly outcomes clearly grow with the size of the control mean. This pattern has been seen in one-way interactions within individual studies; Boston and Chicago had significantly bigger violent-crime effects for those with prior records than those without (Modestino, 2019; Davis and Heller, 2020), and those with prior arrests have larger point estimates for arrests and convictions than those without (Kessler et al., 2021). The analysis here shows that a similar pattern holds across outcomes in the same place (Panel A of Figure 2), across outcomes in different places and times (Panel B), and across subgroups in different places and times (Panel C). Since these estimates were selected based on statistical significance, the pattern does not mean that all youth in groups with higher prevelance of negative outcomes respond more to the treatment; there are other subgroups and outcomes with high control means where there were no significant program impacts. But it does mean that when SYEPs change outcomes, those changes are bigger when the outcomes are more prevalent.

There is a mirror image of the pattern for the few adverse effects as well; youth in groups where the outcome is more common have a bigger adverse response as well. These outcomes are generally much less socially costly than the outcomes that are falling (property and drug crimes sometimes increase, and it is not clear whether the increase in male fertility is from sexual behavior or increased willingness for fathers to be listed on the birth certificate). So while it is worth considering how careful targeting and program adjustments may help to minimize those increases, the overall declines in outcomes like mortality, violent crime, and child protective services likely dominate any cost-benefit calculation.

relative to the control mean. While it is certainly 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 mothers (who have a negative but not significant point estimate). The 2018 cohort of WorkReady consistently faces a floor effect across many outcomes; their program impacts are often significantly more positive than the 2017 cohort, because their baseline rates are so low that there was no room for a decline. This is consistent with the overall lesson from this analysis: that targeting youth at higher risk of these outcomes will generate larger effects.

7.3 Interpreting effect heterogeneity by risk level

Overall, the data seem strikingly consistent with Panel B of Figure 1, suggesting we could extrapolate the absolute magnitude of program effects in new settings as a proportional function of the anticipated control mean. The consistency of this relationship across outcomes could have a number of explanations. First, it could be that a similar mechanism is driving SYEPs' behavioral effects across all these outcomes – a range of crime types, measures of individual behavioral health and mortality, and measures of family stability. Income, changes in beliefs about the future, shifts in time use, or the development of social and self-regulation skills could be similar inputs into the production of these outcomes, generating this kind of proportional shift in multiple outcomes. But there are also alternative explanations. Suppose, for example, that the key behavioral mechanism stems from the interactions between youth and adult program providers, and providers allocate time and attention to youth who are struggling the most. Or suppose there are diminishing marginal returns to adult interaction, such that youth with lower $\mathbf{Y_0}s$ have already benefitted from other adult attention and thus have lower benefits from program-driven investment. Either explanation could generate the pattern of results.

Another possibility is that the overall prevalence of these outcomes is low enough in all these groups that a floor effect is still binding. Even in the point farthest to the right of the graph — the decline in arrests for other crime among those with a prior arrest in the OSC+ 2015 sample — there are 30 arrests per 100 youth in the control group. Even if each of those arrests belonged to a different person, that still leaves 70 control youth for whom the program can not move other-crime arrests below 0. So even at the extreme of the data, it is possible that the graph reflects the sloped portion of Panel A in Figure 1.

Regardless of the reason, the linear relationship between risk and responsiveness holds a crucial lesson for targeting SYEPs. Across all experimental studies of these programs, groups with higher baseline rates of affected outcomes have larger beneficial program effects. There does not seem to be a margin past which youth respond less. To the extent policymakers want to generate declines in the kinds of outcomes measured here, finding ways to recruit and retain youth at elevated risk of any of the focal outcomes—ideally while finding ways to

minimize any adverse effects—is likely to maximize the net social benefits from the program. The risk-response relationship is also good news for policymakers concerned with equity. Since those at the highest risk of harmful outcomes seem to benefit the most, targeting the program to have the biggest impact is equivalent to serving the population that would otherwise be the most disadvantaged.

One concern about this targeting strategy would be if program composition plays a key role in behavior change. In a world where peer interactions are a key input into program effects, a targeting strategy that dramatically alters to whom youth are exposed during the program could change the program's impact. It is perhaps informative, though, that the studies included in Panel B of Figure 2 vary quite a bit in the composition of peer groups within the program. For example, OSC+ 2013 purposefully focused on recruiting a large number of youth at elevated risk of criminal justice involvement, with almost half entering the program with an arrest record. In NYC, on the other hand, only about 3 percent had been arrested at baseline. So at least within the variants of SYEPs that have been experimentally evaluated, being careful not to extrapolate too far out of sample, the results here suggest that targeting populations at higher risk of bad outcomes will increase program benefits. The more likely applicants are to engage in the kinds of risky behavior the programs reduce, the bigger the social benefits from reducing those outcomes are likely to be.

Policymakers may have multiple goals when deciding whom to target with SYEPs, some of which are better served by enrolling youth at lower risk of socially costly outcomes. Providing widespread income transfers, for example, or developing the kinds of skills and connections that help in the labor force may imply different targeting goals (e.g., Davis and Heller (2020) find evidence that employment effects are larger for younger youth more attached to school and less involved in the criminal justice system[34]). But given the high social costs of the type of program effects documented in this paper — crime, incarceration, and the need for child protective services — increasing effects on these outcomes should help SYEPs generate benefits that exceed program costs. And since the subgroups that seem most responsive are also marginalized in other ways, such targeting may help advance social

[34]These more-responsive youth also have higher employment rates in the control group. So this is consistent with the main pattern above of bigger responses for higher control means. But it has different implications for outcomes like employment that have social benefits rather than social costs.

justice and equity concerns as well.

8 Conclusion

Variation in implementation, program structure, who participates, and their responsiveness can sometimes dramatically change an intervention's effects across settings. Understanding that variation should be an important input into decisions about public investments, since it is crucial to predicting whether an expanded investment in one place is likely to replicate the success of any given intervention strategy. This paper assesses and unpacks scale and replicability for one promising type of intervention, SYEPs. These programs consistently reduce criminal justice involvement in the first year after random assignment, and may have some lasting effects as well. They may also help reduce the need for child protective and behavioral health services, although these results are less precise and concentrated among some subgroups.

The insensitivity of the decline in criminal justice involvement across time, location, and program implementation suggests that the basic structure of the program is more important than the details. In support of this idea, there is no detectable heterogeneity in treatment effects across different providers that vary in program type, job placements, staff, and experience with the program. The Chicago results by treatment arm hint at the fact that the added supports of SYEPs may matter above and beyond what is provided in a typical, non-program summer job, but the hypothesis should be tested in a higher-power setting.

Although treatment heterogeneity does not seem related to program structure or delivery, it is related to participant risk level. Youth at higher risk of socially costly outcomes have larger treatment effects. This is relevant for scaling: If programs get so big that the risk level of the population served drops, the program effects are likely to persist but get smaller, as in Philadelphia. But when programs are not universal and make purposeful targeting choices, growing while remaining smaller than the population of those who could feasibly benefit, scaling up without major differences in youth populations is feasible, as seen in OSC+.

The heterogeneity results also suggest something important about the structure of the underlying behavioral response — that at least in the contexts that have been tested so far, there is no such thing as "too late" to generate change. For the outcomes that respond to

treatment, there does not appear to be a margin past which youth fail to respond; rather, making eligibility and targeting decisions that encourage youth at higher risk of crime, family instability, and health problems to participate is likely to generate bigger social gains.

There are limitations to this targeting recommendation. It is possible that massive shifts in program populations, beyond what has already been tried, could change peer exposure in a way that diminishes program impacts. There are also a range of other issues policymakers need to consider. For example, the increased costs of serving more disconnected youth could get high enough to outweigh the increased benefits. At the same time, the benefits of serving youth at very low risk of the outcomes measured here could be low enough that they do not justify program costs; that depends on the size of other benefits not captured by the SYEP studies. Alongside the lesson from this paper that the magnitude of benefits is likely to grow with the prevalence of the outcome in a particular group, detailed consideration of the different social costs across outcomes should inform final decisions about targeting.

Other limitations of the analyses here generate directions for future work. The compliance issues significantly limited statistical power, such that further research on how different program elements matter and how behavioral health outcomes respond would be valuable. The use of administrative data allows large-scale studies like this one, but it also means outcome measures combine changes in the underlying behavior of interest with potential changes in reporting, willingness to seek treatment, or getting caught. Supplementing administrative data with large-scale surveys could help future research parse out these changes.

Despite their limitations, the overall message of the experiments reported here is fairly optimistic. The evidence suggests that SYEPs are not just promising in a way that is "ready for scale," but that they are actually scalable. Their effects seem to be relatively insensitive to context. And the deeper understanding of treatment heterogeneity provided here means that policymakers have additional information about how to replicate and scale strategically to maximize social impact. Unpacking the elements of how interventions are likely to translate into new contexts in this way has the potential to inform public spending decisions more effectively than just establishing which novel interventions "work."

References

Abadie, Alberto, Chingos, Matthew M., and West, Martin R. (2018). "Endogenous Stratification in Randomized Experiments". *Review of Economics and Statistics* 100(4), pp. 567–580.

Aizer, Anna and Doyle, Joseph J. (2015). "Juvenile Incarceration, Human Capital and Future Crime: Evidence from Randomly-Assigned Judges". *The Quarterly Journal of Economics*, pp. 759–804.

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

Athey, Susan and Imbens, Guido (2017). "The Econometrics of Randomized Experiments". In: *Handbook of Field Experiments*. Ed. by Abhijit Binayak Banerjee and Esther Duflo. North-Holland.

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.

Berk, R. A., Lenihan, K. J., and Rossi, P. H. (1980). "Crime and Poverty - Some Experimental Evidence from Ex-Offenders". *American Sociological Review* 45(5), pp. 766–786.

Bloom, Dan (2010). "Transitional Jobs: Background, Program Models, and Evaluation Evidence".

Bloom, Howard S., Orr, Larry L., Bell, Stephen H., Cave, George, Doolittle, Fred, Lin, Winston, and Bos, Johannes M. (1997). "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* 32(3), pp. 549–576.

Card, David, Kluve, Jochen, and Weber, Andrea (2017). "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations". *Journal of the European Economic Association* 16(3), pp. 894–931.

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

Charles, Kerwin Kofi and Luoh, Ming Ching (2010). "Male Incarceration, the Marriage Market, and Female Outcomes". *The Review of Economics and Statistics* 92(3), pp. 614–627.

Davis, Jonathan M.V., Guryan, Jonathan, Hallberg, Kelly, and Ludwig, Jens (2017). "The Economics of Scale-Up". *NBER Working Paper No. 23925*.

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.

Dobbie, Will, Goldin, Jacob, and Yang, Crystal S. (2018). "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges". *American Economic Review* 108(2), pp. 201–240.

Doleac, Jennifer L., Temple, Chelsea, Pritchard, David, and Roberts, Adam (2020). "Which prisoner reentry programs work? Replicating and extending analyses of three RCTs". *International Review of Law and Economics* 62.

Fein, David and Hamadyk, Jill (2018). "Bridging the Opportunity Divide for Low-Income Youth:Implementation and Early Impacts of the Year Up Program". *Pathways for Advancing Careers and Education (PACE)*.

Fisher, Ronald Aylmer (1935). *The Design of Experiments*. Oliver & Boyd.

Friedlander, Daniel (1988). "Subgroup Impacts and Performance Indicators for Selected Welfare Employment Programs". *MDRC*.

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.

Gleicher, Lily (2017). "Juvenile justice in Illinois, 2015". *Illinois Criminal Justice Information Authority*.

Goncalves, Felipe and Mello, Steven (forthcoming). "A Few Bad Apples? Racial Bias in Policing". *Conditionally Accepted at American Economic Review*.

Heckman, James J., Smith, Jeffrey, and Clements, Nancy (1997). "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts". *The Review of Economic Studies* 64(4), pp. 487–535.

Heinrich, Carolyn J., Mueser, Peter R., Troske, Kenneth R., Jeon, Kyung-Seong, and Kahvecioglu, Daver C. (2013). "Do Public Employment and Training Programs Work?" *IZA Journal of Labor Economics* 2.

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

Heller, Sara B. and Bhanot, Syon (in progress). "Overcoming Application and Take-Up Barriers for Summer Youth Employment Programs".

Hinton, Elizabeth, Henderson, LaShae, and Reed, Cindy (2018). "An Unjust Burden: The Disparate Treatment of Black Americans in the Criminal Justice System". *Vera Institute of Justice*.

Holzer, Harry J., Raphael, Steven, and Stoll, Michael A. (2006). "Perceived Criminality, Criminal Background Checks, and the Racial Hiring Practices of Employers". *The Journal of Law and Economics* 49(2), pp. 541–80.

Jepsen, Christopher and Rivkin, Steven (2009). "Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size". *The Journal of Human Resources* 44(1).1, pp. 223–250.

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*.

Kling, Jeffrey R., Liebman, Jeffrey B., and Katz, Lawrence F. (2007). "Experimental Analysis of Neighborhood Effects". *Econometrica* 75, pp. 83–119.

Leos-Urbel, Jacob (2014). "What is a Summer Job Worth? The Impact of Summer Youth Employment on Academic Outcomes". *Journal of Policy Analysis and Management* 33, pp. 891–911.

MDRC (1980). "Summary and Findings of the National Supported Work Demonstration".

MHA Labs (2015). "One Summer Chicago Annual Report 2014".

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

Modestino, Alicia Sasser and Paulsen, Richard (2019). "School's Out: How Summer Youth Employment Programs Impact Academic Outcomes". *Northeastern University Working Paper*.

Mueller-Smith, Michael (2015). "The Criminal and Labor Market Impacts of Incarceration". *University of Michigan Working Paper*.

Ridgeway, Greg and MacDonald, John (2009). "Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops". *Journal of the American Statistical Association* 104(486), pp. 661–668.

Roder, Anne and Elliott, Mark (2020). "Stepping Up: Interim Findings on JVS Boston's English for Advancement Show Large Earnings Gains". *Economic Mobility Corporation*.

Schaberg, Kelsey and Greenberg, David H. (2020). "Long-Term Effects of a Sectoral Advancement Strategy: Costs, Benefits, and Impacts from the WorkAdvance Demonstration". *MDRC*.

Schwartz, Amy Ellen, Leos-Urbel, Jacob, and Wiswall, Matthew (2015). "Making Summer Matter: The Impact of Youth Employment on Academic Performance". *NBER Working Paper No. 21470*.

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

9 Tables

Table 1: WorkReady and OSC+ Descriptive Statistics and Baseline Balance

	WorkReady			One Summer Chicago+		
	Control Mean	Treatment Mean	Test of Difference	Control Mean	Treatment Mean	Test of Difference
N	2,711	1,786	4,497	2,911	2,494	5,405
Demographics						
Age	15.6	15.7	0.87	17.4	17.4	0.82
Male	0.39	0.40	0.64	0.41	0.39	0.20
Black	0.79	0.77	0.36	0.74	0.75	0.74
Hispanic	0.12	0.12	0.63	0.23	0.22	0.93
White	0.04	0.05	0.55	0.01	0.01	0.75
Other Race	0.05	0.06	0.58	0.02	0.02	0.66
Is a Parent	0.015	0.017	0.07			
Contact with Justice System						
Ever Incarcerated as a Juvenile	0.023	0.018	0.31			
Ever Received Juvenile Justice Services	0.024	0.019	0.37			
Ever Arrested	0.046	0.039	0.25	0.223	0.225	0.89
Total Number of Prior Arrests	0.060	0.050	0.23	0.617	0.628	0.88
Violent	0.033	0.019	0.01	0.186	0.186	0.91
Property	0.017	0.018	0.98	0.095	0.097	0.78
Drug	0.003	0.002	0.98	0.085	0.067	0.14
Other	0.007	0.010	0.40	0.251	0.278	0.33
Education						
Enrolled in School	0.87	0.87	0.99	0.74	0.75	0.71
Graduated	0.06	0.08	0.25	0.24	0.24	0.95
Grade	9.5	9.6	0.76	10.7	10.7	0.85
Days Absent	18.1	17.6	0.55	23.9	24.2	0.44
Grade Point Average	2.41	2.46	0.29	2.39	2.40	0.67
Ever Suspended	0.16	0.14	0.38	0.20	0.19	0.31
Receipt of Social Services						
Ever Received Child Protective Services	0.15	0.13	0.07			
Ever Stayed in a Shelter	0.05	0.03	0.01			
Any Behavioral Health Service	0.27	0.25	0.30			
Any Substance Abuse Services	0.01	0.01	0.86			
Any Mental Health Services	0.27	0.25	0.33			
P-value on F-test of treatment-control comparison for all non-behavioral health baseline characteristics			0.55			0.61
P-value on F-test of treatment-control comparison for all behavioral health baseline characteristics			0.84			

Note: WorkReady N=4497, OSC+ N=5405. Enrollment and graduation reported for those with non-missing school records (WorkReady N=4144, OSC+ N=5380). Grade level from application data in WorkReady (N = 4482). Other education measures reported for non-missing data on non-graduates only, excluding charters in Philadelphia (for days absent, WorkReady N=2336, OSC+ N=5308; for suspensions, WorkReady N=2337, OSC+ N=5308; for GPA, WorkReady N=2228, OSC+ N=5158). The Philadelphia school year has 180 days; Chicago has 178. The Test of 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 2: Program Impacts in the First Year After Randomization

	ITT	CM	LATE	CCM
	WorkReady			
Any Juvenile Incarceration	-0.005*	0.014	-0.015*	0.019
	(0.003)		(0.009)	
Any Receipt of Juvenile Justice Services	-0.002	0.013	-0.006	0.018
	(0.003)		(0.009)	
Total Number of Arrests	-0.010**	0.028	-0.030**	0.046
	(0.005)		(0.015)	
Number of Violent Arrests	-0.002	0.011	-0.006	0.016
	(0.004)		(0.010)	
Number of Property Arrests	-0.002	0.008	-0.006	0.012
	(0.003)		(0.008)	
Number of Drug Arrests	-0.003	0.004	-0.007	0.007
	(0.002)		(0.005)	
Number of Other Arrests	-0.004***	0.004	-0.011***	0.010
	(0.001)		(0.004)	
Graduated or Still Enrolled in School	-0.005	0.945	-0.013	0.971
	(0.007)		(0.019)	
	OSC+			
Total Number of Arrests	-0.023	0.176	-0.087	0.166
	(0.015)		(0.057)	
Number of Violent Arrests	0.005	0.037	0.021	0.012
	(0.006)		(0.022)	
Number of Property Arrests	0.000	0.020	0.002	0.019
	(0.005)		(0.018)	
Number of Drug Arrests	-0.012**	0.034	-0.046**	0.052
	(0.006)		(0.022)	
Number of Other Arrests	-0.017*	0.084	-0.063*	0.082
	(0.009)		(0.035)	
Graduated or Still Enrolled in School	-0.001	0.951	-0.003	0.978
	(0.006)		(0.023)	

Note: WorkReady N=4497, OSC+ N=5405. 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. Graduated/still enrolled excludes pre-program graduates and those unmatched to education records; WorkReady N= 3858, OSC+ N= 4077. 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 3: WorkReady Family and Health Impacts in the First Year After Randomization

	ITT	CM	LATE	CCM
Becomes A Parent	0.001	0.011	0.002	0.000
	(0.003)		(0.009)	
Any Receipt of Child Protective Services	-0.006*	0.018	-0.016*	0.013
	(0.003)		(0.010)	
Any Stay in a Shelter	0.002	0.001	0.006	0.000
	(0.002)		(0.005)	
Any Receipt of Behavioral Health Services	-0.011	0.113	-0.033	0.108
	(0.009)		(0.026)	
Any Receipt of Substance Abuse Services	-0.003	0.008	-0.008	0.009
	(0.003)		(0.008)	
Any Receipt of Mental Health Services	-0.010	0.110	-0.029	0.104
	(0.009)		(0.026)	

Note: N=4497. 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. 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 *p<0.1, **p<0.05, ***p<0.01

Table 4: OSC+ Outcomes in the First Year After Randomization by Treatment Arm

	ITT	CM	ITT	CM	Test of
	Job Only		Job and Mentor		Difference
Total Number of Arrests	-0.015	0.174	-0.031*	0.177	0.424
	(0.018)		(0.018)		
Number of Violent Arrests	0.004	0.044	0.007	0.031	0.692
	(0.007)		(0.008)		
Number of Property Arrests	0.002	0.014	-0.001	0.026	0.605
	(0.006)		(0.006)		
Number of Drug Arrests	-0.007	0.037	-0.017**	0.030	0.146
	(0.007)		(0.007)		
Number of Other Arrests	-0.014	0.079	-0.020*	0.090	0.609
	(0.011)		(0.011)		
Graduated or Still Enrolled in School	-0.005	0.947	0.004	0.954	0.296
	(0.008)		(0.008)		

Note: N=5405. Graduated or still enrolled in school excludes pre-program graduates and those unmatched to education records; N= 4077. Table shows estimated intent-to-treat (ITT), controlling for baseline covariates and randomization block. CM is control mean; CCM is control complier mean, rounded to 0 when estimate is negative. 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 5: Treatment Effect on Combined Index by Predicted Risk Level

Panel A: Intent to Treat Effect, Index		
	WorkReady	OSC+
Full Sample	-0.065***	-0.051**
	(0.025)	(0.021)

Panel B: WorkReady Intent to Treat Effect by Predicted Risk Level

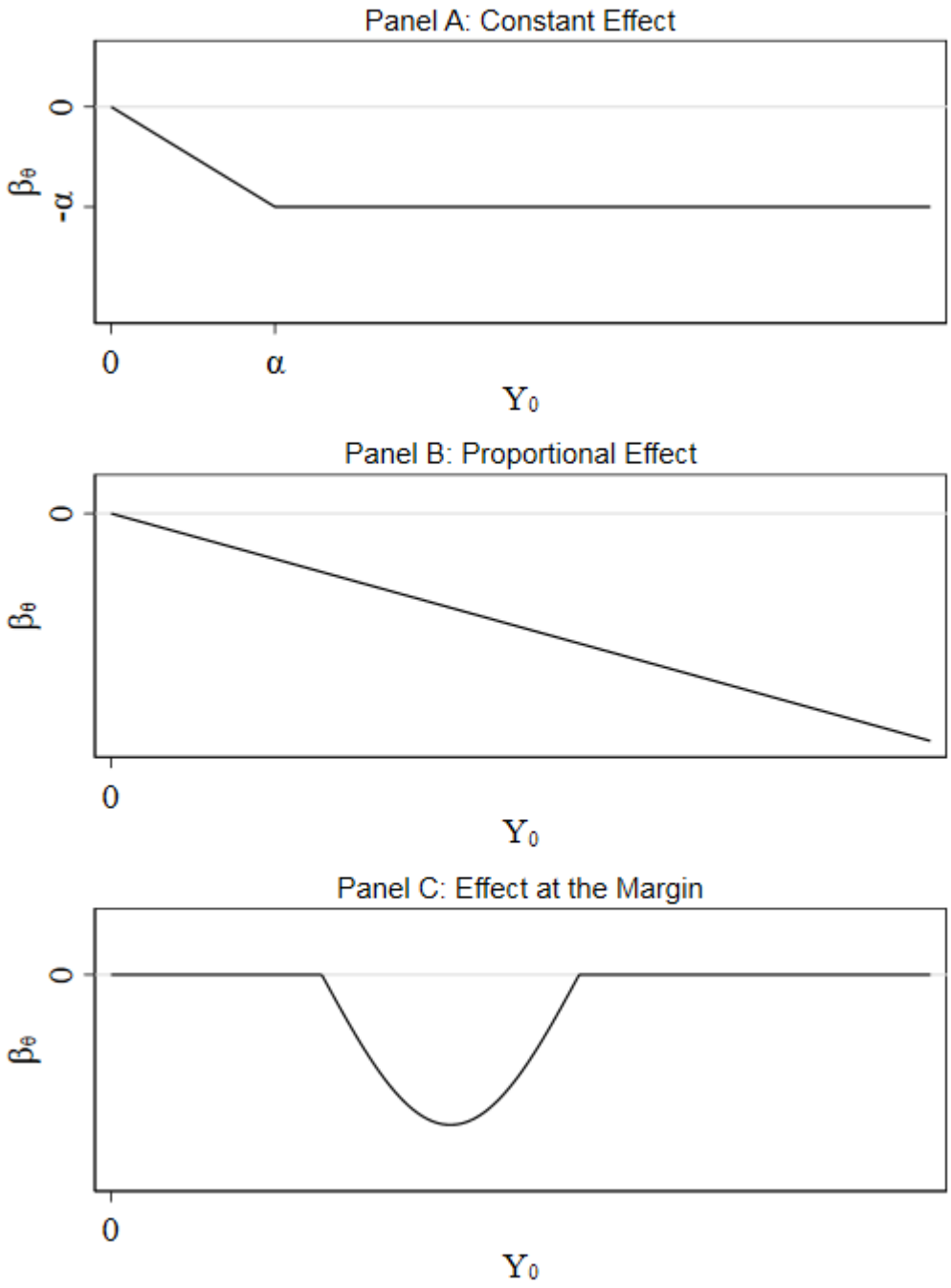
Predicted Risk Level	Repeated Split Sample	Leave One Out	CM	First Stage
Low	-0.007	0.018	-0.169	0.405
	(0.016)	(0.020)		
Medium	-0.020	-0.078*	-0.059	0.328
	(0.022)	(0.041)		
High	-0.169**	-0.190**	0.207	0.277
	(0.066)	(0.081)		

Panel C: OSC+ Intent to Treat Effect by Predicted Risk Level

Predicted Risk Level	Repeated Split Sample	Leave One Out	CM	First Stage
Low	-0.008	-0.005	-0.185	0.281
	(0.008)	(0.011)		
Medium	-0.018	-0.012	-0.136	0.258
	(0.013)	(0.019)		
High	-0.115**	-0.132**	0.321	0.262
	(0.054)	(0.055)		

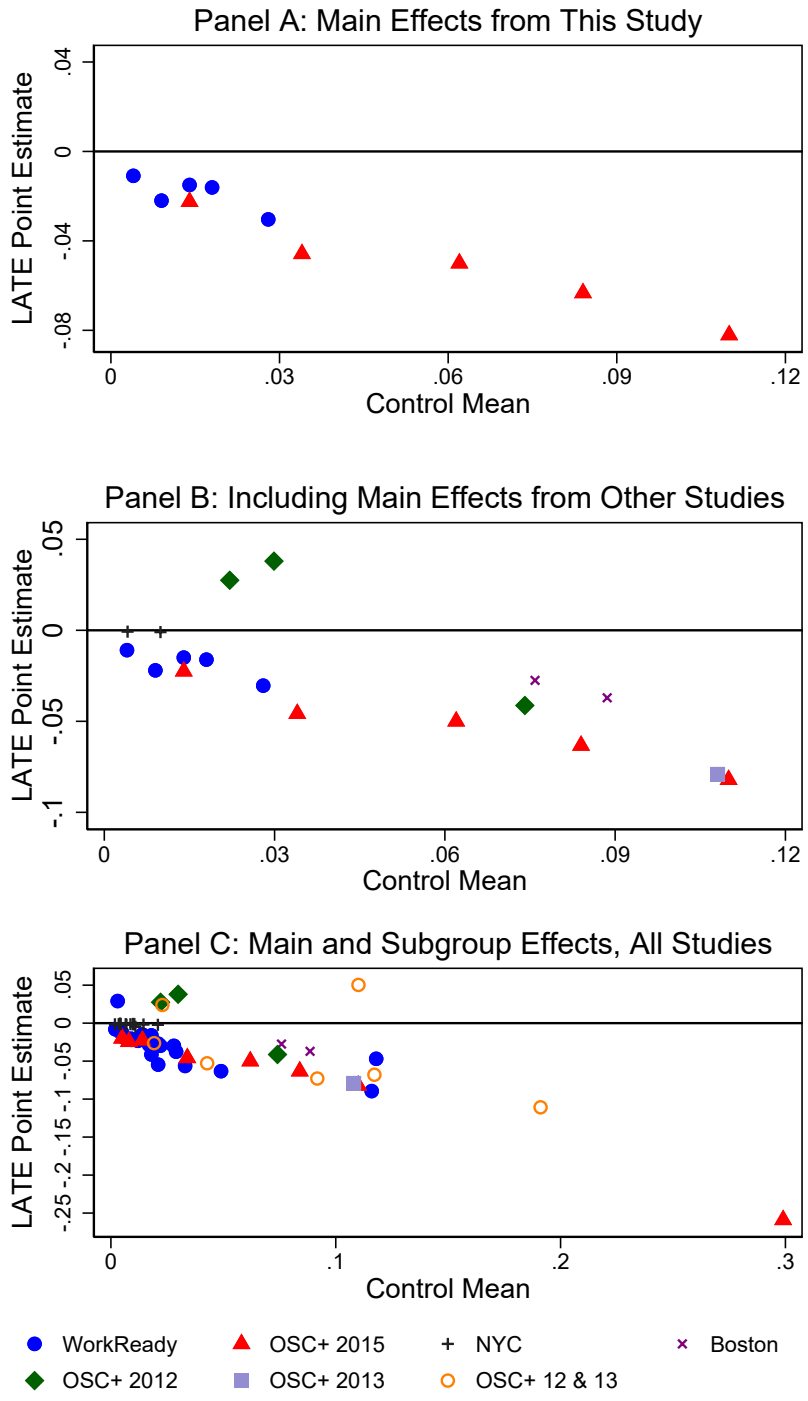
Note: WorkReady N=4497, OSC+ N=5405. Estimation from the Abadie, Chingos, and West (2018) procedure. CM is the control mean for each group, assigned in the leave one out estimation. Table shows the effect of each program on a standardized index of the outcomes that have significant changes in the main estimates: incarceration, total arrests, and receipt of child protective services (WorkReady) and drug and other arrests (OSC+). 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 1: Stylized Treatment Effects Relative to Counterfactual Outcomes with Different Types of Heterogeneity



Note: Figure shows theoretical shape of risk-responsiveness relationship under different types of treatment heterogeneity. See Section 7.1 for discussion.

Figure 2: Size of LATEs 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.