Journal of Public Economics 209 (2022) 104617

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

When scale and replication work: Learning from summer youth employment experiments

Sara B. Heller *

University of Michigan, United States
NBER, United States

ARTICLE INFO

Article history:
Received 7 July 2021
Revised 28 December 2021
Accepted 29 January 2022
Available online 4 April 2022

Keywords:
Summer youth employment
Treatment heterogeneity
Replication
Scale
Crime prevention

ABSTRACT

This paper combines two new summer youth employment experiments in Chicago and Philadelphia with previously published evidence to show how repeated study of an intervention as it scales and changes contexts can guide decisions about public investment. Two sources of treatment heterogeneity can undermine the scale-up and replication of successful human capital interventions: variation in the treatment itself and in individual responsiveness. Results show that these programs generate consistently large proportional decreases in criminal justice involvement, even as administrators recruit additional youth, hire new local providers, find more job placements, and vary the content of their programs. Using both endogeneous stratification within cities and variation in 62 new and existing point estimates across cities uncovers a key pattern of individual responsiveness: impacts grow linearly with the risk of socially costly behavior each person faces. Identifying more interventions that combine this pattern of robustness to treatment variation with bigger effects for the most disconnected could aid efforts to reduce social inequality efficiently.

1. Introduction

As policymakers consider how to reduce poverty, 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 settings but had different effects when scaled up or moved elsewhere.[1] Anticipating which interventions will successfully replicate and scale requires understanding whether two main aspects of program growth or context generate heterogeneous treatment effects: (1) variation in what the treatment is (e.g., program structure, staff quality, and counterfactual opportunities), and (2) variation in who is served (see Davis et al., 2017; Al-Ubaydli et al., 2020, for broader discussions of the challenges of scale).

This paper shows how repeated study of an intervention as it scales and changes contexts, in this case summer youth employment programs (SYEPs), can guide decisions about public investment. SYEPs are a useful demonstration of the uncertainty involved with expanding and replicating "evidence-based" programs. Experiments in Chicago, New York, and Boston have found generally similar patterns of SYEPs' effects: large declines in criminal justice involvement and violence, despite little improvement in future employment on average (Davis and Heller, 2020; Gelber et al., 2016; Heller, 2014; Kessler et al., 2021; Modestino, 2019). Education impacts are more mixed, with most studies finding small or no improvements in high school or college outcomes (Davis and Heller; Gelber et al., 2016; Heller, 2014; Leos-Urbel, 2014; Schwartz et al., 2021), and one showing larger benefits (Modestino and Paulsen, 2022). But there is reason to worry that efforts to expand on SYEPs' success might diminish their effectiveness. The Chicago and Boston studies focus on a relatively small subset of the cities' summer programs, with service providers selected into each evaluation. In other settings, unrepresentative provider selection and difficulty with staffing and implementation quality as programs scale (e.g., Allcott, 2015; Bhatt et al., 2021; Jepsen and Rivkin, 2009) have made the effects of expansion

* Address: University of Michigan, Department of Economics, 611 Tappan Ave, 238 Lorch Hall, Ann Arbor, MI 48109, United States.
E-mail address: sbheller@umich.edu

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

https://doi.org/10.1016/j.jpubeco.2022.104617
0047-2727/

attemps smaller than in initial studies. Evidence from NYC's SYEP might increase confidence, since it 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 shifts in residential population or local conditions have changed the impact the SYEP would have today. To make informed decisions about where future investments are likely to succeed, policymakers require more than past successes; they need a better understanding of how net effects respond to variation in what the treatment is and whom it serves.

To help provide this understanding, this paper combines two new randomized controlled trials with evidence from prior SYEP experiments. In addition to providing a valuable replication exercise, the two new RCTs demonstrate what changes as program administrators recruit additional youth, hire new local providers, find more job placements, and vary the content of their programs. The first new experiment tests a program called One Summer Chicago Plus (OSC+) in summer 2015, which tripled in size (n = 5,405) relative to the 2012 and 2013 versions of the program studied elsewhere (Davis and Heller, 2020; Heller, 2014). OSC+, which targets youth at high risk of violence, scaled up without generating much change in its applicant or complier population. It also grew from 5 to 19 providers and changed the youth development programming it provided over time. The second new experiment studies Philadelphia's WorkReady program, which has not been evaluated before, in the summers of 2017 and 2018 (n = 4,497, a subset of the larger program).[2] WorkReady's universal, city-wide scale means it served a less targeted, less criminally active population than OSC+. It also involved much broader variation in program models across 50–60 providers, with little scope for providers to select into or out of the evaluation.

In many settings, this kind of variation in programming, providers, and populations has undermined programs' ability to replicate originally positive results. But for these SYEPs, results show that both new experiments generated large reductions in criminal justice contact in the first year after random assignment, with declines in some types of arrests and incarceration on the order of 50–80 percent. For cohorts where enough time has passed to measure longer-term effects, there were also indications of arrest declines during years 2 and 3.[3] Both experimental designs involved randomization to either different program models or different providers. These tests find no significant variation by model or provider, which could help to explain why effects replicate across different scales and contexts. Neither experiment is fully powered to test this question, however, in part because non-compliance resulted in first stages that are both around 0.3. Still, all existing SYEP experiments show relatively similar proportional declines in criminal justice involvement across considerable variation in the number of providers and jobs, as well as the type of enrichment activities they offer. And all show crime declines with similar time paths that rule out a simple incapacitation mechanism, with effects continuing to accrue after the end of the program. This replicability, at least across the large urban settings where SYEPs have been tested, is somewhat unusual in human capital development programs. It suggests that something about the core program structure, not a specific implementation, setting, or population, drives behavioral change.

The fact that proportional declines are relatively constant across different populations implies that the absolute number of crimes prevented is larger for groups with higher control means. With this in mind, the second part of the paper turns to how treatment effects (β) vary with the population served. I lay out a framework motivating why heterogeneous responses across youth who face different counterfactual risk levels (Y_0) is of particular interest for questions of scale and replication. In addition to informing optimal targeting, the distribution of gains by counterfactual risk level also informs questions of equity 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 β may also reflect 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.

I use two strategies—endogenous stratification within the two new experiments and an analysis of variation in youths' risk level across all published SYEP experiments—to estimate the relationship between Y_0 and β. First, within each new experiment, I generate an index of all the socially costly outcomes with significant treatment effects. I 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 13–26 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 shape of the relationship between risk group and response.

The second strategy brings cross-study variation to the estimation of the risk-responsiveness relationship. I collect 62 statistically significant impact estimates across the two experiments here and four other peer-reviewed experimental studies of SYEPs. Across locations, outcomes, and subgroups, I 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.[4]

Whether the correlation between the risk of harmful outcomes individuals face and their responsiveness is causal or a result of other factors that regularly vary with Y_0 across settings, its consistency provides guidance for how effects may change in new contexts or expanded programs: magnitudes are likely to scale with the risk level of the population served. The fact that effects grow with Y_0, at least among the populations who have been part of prior studies, provides some evidence against the idea that there is a point when it is "too late" to change behavior. And the fact that initially worse-off youth benefit the most suggests a virtuous complementarity between efficiency and equity in targeting decisions.

[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. I nonetheless perform similar multiple testing adjustments, described in the methods section below.

[3] The main text focuses on criminal justice impacts, since those are the outcomes most consistently measured across settings, providing the most opportunity to analyze impact heterogeneity across studies. The appendix also reports education impacts, which are generally null but suffer from considerable missing data due to high charter school enrollment in Philadelphia. Family and health measures from social service records available only in Philadelphia, which have not been measured elsewhere, are also in the appendix. There are promising indications that participation may decrease the need for child protective services, and perhaps substance abuse and mental health services, especially among boys and Black youth. But estimates are too imprecise to survive adjustments for multiple hypothesis testing, and some outcomes are quite rare in the sample. Because these were pre-specified outcomes, I include these results in the individual treatment heterogeneity analysis that follows. But the imprecision means that confirming SYEPs' impacts on family and individual health and well-being should be a priority for future work.

[4] This pattern is not driven by differences in take-up rates; the relationship between intent-to-treat effects and control means looks very similar. Nor is it a by-product of selecting statistically significant outcomes; the pattern remains when accounting for each estimate's variance and when including all main effects.

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 findings suggest that shifting SYEPs' focus towards populations at elevated risk of crime and related negative outcomes could help maximize program benefits on those outcomes, and would likely do so across different contexts and program designs. Identifying other interventions that, like SYEPs, seem to be robust to substantial variation in implementation while also generating the largest benefits among those facing the most challenges, could help efforts to reduce social inequality efficiently. More broadly, 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

This section provides a brief overview of each program and experimental design, focusing on the comparisons within and across studies that contribute to our understanding of scale and treatment heterogeneity. All programs serve teenagers and young adults, and all share the basic elements of job readiness training, a 6–8 week part-time job placement at or near minimum wage, and some type of professional or personal development activities. Table 1 provides additional details about program elements for the two new experiments reported here, as well as other studies that contribute to the cross-study comparisons in Section 7. For more discussion of program and study design, see Appendix A.

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. Two key sets of changes to the 2015 program facilitate the study of what happens as programs scale and adapt over time.[5] First, the program almost tripled in size relative to the first 2012 cohort, from 700 to 2,000 program slots. In addition to requiring program staff to identify 3 times as many job placements, the scale-up also required hiring almost 4 times as many local providers to implement the program (19 in 2015 compared to 5 in 2012) and recruiting students at almost 4 times as many schools (49 compared to 13). This growth helps to identify whether a program that purposefully targets young people at elevated risk of violence can maintain the same youth population as it scales, and whether the challenges of recruiting new providers and finding new jobs at a much larger scale diminish treatment effects.

The second set of changes helps to identify whether treatment effects are driven by specific elements of programming. Unlike in the original studies, which included a social-emotional learning curriculum and a separate, paid adult mentor, program operators tested two versions of the 2015 program. Half the participants worked in their jobs for 25 hours per week, with no separate adult mentor and no additional youth development curriculum.[6] The other half worked for 20 hours per week, and for 5 hours on Fridays engaged in a Civic Leadership Foundation curriculum focused on civic leadership.[7]

A total of 5,405 youth applied to OSC+. The research team grouped them by geography to help minimize commute times, then randomly assigned each individual to the control group ($n = 2,911$), the job-only group ($n = 1,252$), or the job + mentor group ($n = 1,242$) within strata. All individuals were also randomly assigned to a provider within strata.

2.2. WorkReady

In contrast to the more targeted OSC+, WorkReady is a universal summer jobs program open to all city youth, run by the Philadelphia Youth Network (PYN). Its operation at broader scale—between 8 and 10 thousand slots during the study summers—changes the population served relative to OSC+, generates the need for more job placements, creates considerably more variation in both local provider agencies and program details, and reduces the scope for providers to select into the evaluation based on their expected treatment effect. In the 2017 and 2018 study years, PYN contracted with 50–60 local agencies to provide 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.[8] Professional development activities were left up to providers, so they varied considerably in structure and content across agencies, ranging from developing business models to sexual health education.[9]

In summers 2017 and 2018, a subset of WorkReady applicants entered a randomized lottery to allocate the limited num-

[5] Changes to program details are common in SYEPs; OSC+ has continued to change since 2015. Although there was also a 2013 OSC+ study, I focus here on comparing 2015 to the original 2012 cohort. The 2013 program involved explicitly different eligibility criteria, recruiting only males, some from the criminal justice system, to test for effect heterogeneity. So the 2012 cohort, where recruitment worked like the current study, is the most useful comparison for isolating what happens when the same approach scales. The 2013 estimates contribute to the analysis of treatment heterogeneity in the second half of the paper.

[6] In practice, program providers were quite resistant to removing the additional mentorship, so they replaced the mentors with "adult supervisors" who provided less personalized and intensive support, but who were nonetheless available as needed. While this makes the test of the difference slightly less informative, site observations suggested that there was still a difference in the amount of adult support offered across treatment arms, just less of a difference than was originally intended. Because the program only officially provided mentors to half the sample, the increased scale does less to identify the challenge of recruiting additional mentors as the program grew than it does to identify the challenge of finding new job placements (from 2012 to 2015, mentors were provided to an additional 300 youth). One other program change is worth highlighting, since it complicated recruitment: The program obtained a waiver against the city's minimum wage increase, so youth were still paid $8.25 per hour, compared to a concurrent shift to a $10 minimum wage in the regular labor market.

[7] 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.

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

[9] As part of the study, the research team conducted open-ended interviews with 18 study participants (13 in the treatment group) as well as field observations. One of the core conclusions of that work was how much the experience of the program varied across individuals based on job assignment; timing, content, and instructional details of professional development activities; and relationships with supervisors. One might hypothesize that this kind of treatment variation would be likely to generate effect heterogeneity, especially given other work finding that two-thirds of examined multi-site workforce development and education studies show significant variation in treatment effects across sites, and variation is more likely when program details are not highly codified (Weiss et al., 2017).

Table 1
Comparison of SYEP Programs.

Program	WorkReady 2017/2018	OSC+ 2015	OSC+ 2012	OSC+ 2013	Boston 2015	New York City 2005–2008
Approx. Slots City-Wide	8300/9700	24,000	17,000	20,000	10,000	54,000
Approx. Slots In Study	1100/375	2000	700	1000	1186	54,000
Num. Providers In Study	59/45	19	3	7	1	59
Length	6 weeks	7 weeks	8 weeks	6 weeks	6 weeks	Up to 7 weeks
Hours Per Week	20	25	25	25	25	Up to 25
Hourly Wage (Nominal)	$7.25 to $10.00	$8.25	$8.25	$8.25	$9.00	$6.00 to $7.15
Job type	Government, nonprofits, private sector	Government, nonprofits, private sector, infrastructure	Government, nonprofits	Government, nonprofits, private sector	Government, nonprofits, private sector	Government, nonprofits, private sector
Eligible Population	• 14–21 year olds • All city residents	• 16–21 year olds • 49 high-violence CPS high schools	• 14–21 year olds • 13 high-violence CPS high schools	• 16–22 year olds • Male only • Justice agencies and general OSC applicants	• 14–24 year olds • All city residents	• 14–21 year olds • All city residents
Separate Adult Mentor	No	Randomly assigned to 50% of participants	Yes	Yes	No	No
Training and Enrichment	• Professional development sessions throughout summer	• 1 week job readiness training • 5 h/week civic leadership curriculum randomly assigned to 50% of participants	• 1 day job readiness training • 2 h/day social-emotional curriculum randomly assigned to 50% of participants	• 1 day job readiness training • 2 h/day social-emotional curriculum • Some post-summer activities	• 20 h job readiness and professional development training	• 17.5 h job readiness, career exploration, financial literacy training

Note: First two columns are studies in this paper. The other two OSC+ columns are from Davis and Heller (2020) and Heller (2014). The Boston information is from Modestino (2019). The NYC information is from Gelber et al. (2016).

ber 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 at-scale operation; there was limited scope for contracted agencies to select into or out of the evaluation. The choice margin providers faced was at the point of take-up. To facilitate cooperation among providers, PYN did not force providers to serve treatment youth, and they discouraged but did not prohibit providers from serving control youth (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 (N = 1,554 across 39 providers). Youth with no pre-existing provider relationships could submit online applications directly to PYN (N = 1,838). These latter applicants were stratified by geography and age, then individually randomly assigned within strata to both treatment or control groups and to 20 different providers (again generating random variation in provider assignment within strata). In 2018, there was only one lottery consisting of youth who applied to PYN without connections to a provider (N = 1,105). In this cohort, there was no random variation in provider; PYN matched youth to one of 45 providers. In both study years, I 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. I use application and program participation data from the organizations in charge of administering each program. In both cities, I use administrative police records to measure arrests, and in Philadelphia, service records from the City's integrated data system, known as CARES, to measure juvenile incarceration (including both detention and prison, but not any adult incarceration) and related court-ordered services. The main text focuses on criminal justice outcomes, which are the best measured and most comparable across studies. Appendix B discusses the details of other available education and social service data, which include measures of child protective service receipt, mental health and substance abuse treatment, homeless shelter use, and fertility; it also explains the data linkage process.

Arrest records cover lifetime histories for the Chicago sample but 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. I 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 I assign zero arrests for unmatched youth.[10]

4. Analytical methods

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

$$Y_{ist} = \beta_0 + \beta_1 T_{is} + B_2 X_{ist-1} + \gamma_s + \varepsilon_{ist}$$

where Y_{ist} is the outcome of interest for individual i in randomization strata s in period t. T_{is} is an indicator for individual i being randomly assigned to be offered a program slot. X_{ist-1} is a set of individual i's pre-randomization characteristics, and γ_s is a vector of randomization strata fixed effects (see Appendix F for a list of baseline covariates). For any missing baseline covariates I impute 0s and include an indicator for missingness. I show results with no baseline covariates other than the strata fixed effects (and for OSC+, duplicate application indicators) required for identification as a robustness check in Appendix F. To ease interpretation, the main analysis uses ordinary least squares. Since outcomes are either indicators or counts, Appendix F.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, I use random assignment as an instrument for ever participating, defined as having more than 0 hours recorded in program records. Given the two-sided noncompliance, this estimator is a local average treatment effect (LATE) for compliers rather than a treatment-on-the-treated effect. To assess the magnitude of these effects, I report estimates of control complier means (CCMs) as a baseline.[11]

In addition to reporting heteroskedasticity-robust standard errors, clustered on person in WorkReady where 132 applicants appear in both cohorts, I 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 F.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; Benjamini and Hochberg, 1995; Westfall and

[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 (Goncalves and Mello, 2021; Ridgeway and MacDonald, 2009; 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 to future individual and family outcomes (e.g., Aizer and Doyle, 2015; Charles and Luoh, 2010; Dobbie et al., 2018; Holzer et al., 2006; Mueller-Smith, 2015) means there still a large social benefit to reducing arrests, even if some of the change is not driven by changes in underlying crime.

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

Young, 1993). I 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) and the type of arrests (violent, property, drug, and other in both studies). The appendix also reports impacts and multiple testing adjustments for family outcomes (child protection services, shelter use, and fertility) and behavioral health (substance abuse and mental health services), which are only available in Philadelphia.

To test whether there is significant treatment variation across providers, I use the subset of the data with random variation in provider assignment (all of Chicago and part of the 2017 sample in Philadelphia). I include baseline covariates, strata 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 treatment effect model.[12] Since provider assignment is random within strata, the regression does not require the inclusion of provider fixed effects; however, their inclusion does not change the results. Each stratum with random provider variation contains 2–4 providers.

To test for heterogeneity by risk, I implement both the leave-one-out and repeated-split-sample procedures in (Abadie, Chingos and West, 2018). Details are in Section 7.2.

5. Descriptive statistics and compliance

5.1. Sample composition as programs scale

Table 2 shows baseline characteristics for both the WorkReady and OSC+ study populations prior to random assignment. The table demonstrates three key points, with the third discussed in the next section. First, randomization worked: No more of the differences are significant than would be expected by chance, and as shown in the last two rows, the tests of joint significance in both studies confirm that treatment and control groups are balanced.[13]

Second, by comparing the populations across studies, we can investigate how the applicant population changes with the scale of an SYEP. Relative to the initial 2012 OSC+ study, the 2015 program roughly tripled the number of slots and quadrupled the number of schools in which recruiting occurred. Yet many of the characteristics of the 2015 population—about 40 percent male, 22 percent with an arrest record, GPAs of 2.4, and 24 days absent—are fairly similar to the initial 2012 cohort. The 2012 applicants were a little younger (16.3 rather than 17.4 on average, because 14- and 15-year-olds were eligible in 2012), but also about 40 percent male, 20 percent with an arrest record, 2.4 GPAs, and 33 days absent (Davis and Heller, 2020; Heller, 2014). The one major difference between samples is that the current study participants were about 23 percent Hispanic, compared to 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. The relative 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 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.[14]

By contrast, WorkReady is a larger, less-targeted program open to every young person in Philadelphia. It therefore serves a less disadvantaged population. Applicants were considerably less involved in the criminal justice system (only 4 percent had an arrest record) and slightly more engaged in school (18 days absent and 15 percent with a prior suspension, compared to 24 days absent and 20 percent with suspensions in Chicago).[15] The additional social service data available in Philadelphia highlights that despite the differences that come with universal eligibility, SYEP applicants still face more challenges than the average City youth. About 14 percent of applicants were in families that previously received child protective services (some when they were very young); 4 percent had stayed in a homeless shelter; 1.6 percent were parents themselves; and about 26 percent had received behavioral health services, mostly mental health care. These rates are about 50 percent higher than the population of youth in the City's service database who did not apply to WorkReady.

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 program administrator, PYN. To minimize provider resistance to the new lottery system and increase 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. Combined with the fact that the 2018 study focused solely on an

[12] 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 provider assignment is random, the assumptions for random effects are met by construction; provider assignment is not correlated with any other covariate, conditional on strata. So I focus on the random effect approach to aid power.

[13] 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 exclude variables that are linear combinations of the others. 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.

[14] 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, 2015). The ACS reports a little over 150,000 people between 15 and 17 living in Chicago. So while there is likely 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 (as we see in Philadelphia).

[15] Although the study sample is not perfectly representative of the WorkReady program as a whole (applicants without pre-existing relationships with providers are over-represented, see Appendix A), it does reflect a broad subset of the program, with youth at almost all providers participating.

Table 2
WorkReady and OSC+ Descriptive Statistics and Baseline Balance.

	WorkReady (Philadelphia)			One Summer Chicago Plus		
	Treatment Mean	Control Mean	Complier Mean	Treatment Mean	Control Mean	Complier Mean
N	1,786	2,711		2,494	2,911	
Demographics						
Age	15.7	15.6	15.5	17.4	17.4	17.3
Male	0.40	0.39	0.34	0.39	0.41	0.43
Black	0.77	0.79	0.82	0.75	0.74	0.82
Hispanic	0.12	0.12	0.09	0.22	0.23	0.16
White	0.05	0.04	0.02	0.01	0.01	0.01
Other Race	0.06	0.05	0.07	0.02	0.02	0.01
Is a Parent	0.017	0.015	0.008			
Contact with Justice System						
Any Juvenile Incarceration	0.018	0.023	0.013			
Any Juvenile Justice Services	0.019	0.024	0.024			
Ever Arrested	0.039	0.046	0.039	0.225	0.223	0.217
Number of Prior Arrests	0.050	0.060	0.052	0.628	0.617	0.477
Violent	0.019**	0.033	0.039	0.186	0.186	0.136
Property	0.018	0.017	0.008	0.067	0.085	0.065
Drug	0.002	0.003	0.000	0.097	0.095	0.093
Other	0.010	0.007	0.006	0.278	0.251	0.183
Education						
Enrolled in School	0.87	0.87	0.93	0.75	0.74	0.76
Graduated	0.08	0.06	0.04	0.24	0.24	0.24
Grade	9.6	9.5	9.4	10.7	10.7	10.7
Days Absent	17.6	18.1	12.8	24.2	23.9	24.0
Grade Point Average	2.46	2.41	2.44	2.40	2.39	2.35
Ever Suspended	0.14	0.16	0.16	0.19	0.20	0.22
Receipt of Social Services (Indicators for Each Service Type)						
Child Protection	0.13*	0.15	0.15			
Homeless Shelter	0.03***	0.05	0.05			
Behavioral Health	0.25	0.27	0.26			
Substance Abuse	0.01	0.01	0.01			
Mental Health	0.25	0.27	0.25			
P-value on joint F-test, non-behavioral health			0.55			0.61
P-value on joint F-test, behavioral health			0.84			

Note: WorkReady N = 4497, OSC+ N = 5405. Stars on treatment column indicate p-value from test that treatment and control means are equal, adjusting for randomization block (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$). Compiler means calculated with Abadie (2003) kappa weights, adjusting for stratified randomization. Race/ethnicity coded as mutually exclusive categories based on self-identification from program applications. 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. 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.

applicant pool that did not apply directly to providers (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 = 603).

Given these recruitment processes, complying with randomization is a function of both provider and individual decisions. The third column of Table 2 shows the resulting average complier characteristics for each study.[16] On most characteristics, compliers are not all that different from the applicant population. In both cities, compliers are somewhat more Black than the sample as a whole. In Philadelphia, they are slightly less likely to have been incarcerated and more likely to have been enrolled in school in the prior year. In Chicago, compliers were less involved in the criminal justice system than the full applicant pool. But for the most part, there is not a lot of observable selection into compliance in either setting; the main differences across studies come from the differences in who applied to each program.

6. Results

This section presents SYEP impact estimates for the crime outcomes that are available across both cities. Education and other family and health outcomes, which are less consistently available across the two cities and so less conducive to an analysis of treatment heterogeneity across contexts, are reported in Appendix C.[17] The similarity of proportional changes across the crime estimates reported here, as well as in previous studies, is useful on its own as a replication exercise. It also suggests that variation in what the treatment is—program model, staff capacity and experience, and local context—is less important to generating treatment effects than the basic intervention approach itself.

[16] Complier means are calculated with kappa weights that account for the varying treatment probabilities across strata, per Abadie (2003).

[17] As shown in the appendix, there is no change in school persistence, the best-measured education outcome. There are some suggestions of improvements in family and health outcomes that could be related to income, time use, or personal skills like self-efficacy. Child protective service receipt shows a marginally significant decline, especially among Black youth, and behavioral health services decline among boys. But these results are sensitive to adjustments for multiple testing, so they serve to generate hypotheses for exploration in future work rather than establish clear impacts. Appendix Tables A8 through A11 show no clear improvements in other education outcomes, regardless of how missing data are imputed. If anything, the point estimates tend to be negative; see Appendix F.4 for discussion.

Table 3
Program Impacts in the First Year After Randomization.

	ITT	CM	LATE	CCM
	WorkReady (Philadelphia)			
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
	(0.003)		(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)	
	One Summer Chicago Plus			
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)	

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. Robust standard errors in parentheses, clustered by person for WorkReady, where the same person can appear in both cohorts. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

6.1. Main effects

Table 3 shows the estimated ITT and LATE for criminal justice involvement in the first year after randomization. In both cities, the SYEP 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 incarceration result is marginally significant ($p = 0.08$), 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 G), 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 here previews the risk-responsiveness relationship I investigate below: Outcomes with larger control means also show larger point estimates.

A number of the main results cross traditional significance thresholds after adjustments for multiple testing (see Appendix F.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.

It is worth noting that the type of crimes that respond to SYEPs here differs somewhat from prior studies of OSC+. In prior work (Davis and Heller, 2020; Heller, 2014), OSC+ crime declines were driven by decreases in violent-crime arrests (which is why violence was the primary pre-specified outcome in the pre-analysis plan). While it is possible the shift has to do with 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, despite very similar baseline rates of arrest (1.1 and 3.7 violent-crime arrests per 100 control youth in year 1 here, relative to 7.4 per 100 in the 2012 study and 10.8 per 100 in the 2013 study).[19] Looking at proportional changes, the 30–40 percent declines in violence found in prior studies are within the confidence intervals of both the WorkReady and OSC+ studies here. So although

[18] The drug arrest result in Philadelphia is exactly at the standard significance threshold, $p = 0.100$, which I treat as marginally significant.

[19] 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 a fundamental change in the population served. 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 significant treatment effect on violence is because the lower occurrence and recording of violent events makes behavioral changes harder to detect in the data.

there is not enough precision to confirm a similarly-sized decline in violence in these studies, I can not rule it out. The time path of effects is also similar across studies. Appendix Section D presents evidence that as in prior work, being busy over the summer is not the key mechanism; program effects continue to accrue long after the program ends.

6.2. Variation across program structure and delivery

Because variation in what the treatment actually is can contribute to difficulty with scale and replication, the research design included experimental variation in both program model and local provider. In Chicago, I randomly varied the structure of mentorship and the enrichment curriculum across two treatment arms, and in both cities I generated random variation in provider assignment. In neither case can I reject the null that treatment effects are the same (neither across treatment arms nor across providers). However, in part due to the relatively high non-compliance, these tests are quite underpowered (see discussion and results in Appendix Section E).

Across-study comparisons are also informative about whether mentors, particular enrichment curricula, or variation in provider experience and quality are crucial to program success. As discussed above and shown in Table 1, from 2012 to 2015, OSC+ hired almost 4 times as many providers, introducing more variation in provider experience and characteristics; tripled in size, requiring a big expansion in the number of jobs provided; and changed from mentorship with a social-emotional learning curriculum to either a civics curriculum or no curriculum and no mentor. The Philadelphia study expands the scale even further, with the city-wide programming encompassing broader variation in program models, a wide range of enrichment curricula, no separate paid mentors, and about 3 times as many providers as in even the expanded OSC+.

The consistency of criminal justice involvement declines suggest that neither scaling up nor adding variation in program models and delivery undermines the programs' ability to reduce criminal justice involvement. Of course, this does not rule out the possibility that changes in scale, providers, and program details can generate substantively important treatment heterogeneity; well-powered tests of specific kinds of variation would be useful in future work, and one particular pattern of heterogeneity is discussed more in the next section. But the replications here do demonstrate that none of the factors that varied across studies constitute the "key ingredient" for having a positive impact. Rather, something about the basic program structure that all these SYEPs share is enough to reduce crime.

7. Variation by individual risk level

In the main results, the proportional change in crime outcomes relative to the control group is similar across contexts; that implies the absolute magnitude of the point estimates is larger when control means are larger. Motivated by this pattern, which also holds across a number of subgroup results in the appendix on other socially costly outcomes, this section focuses on estimating individual heterogeneity by risk, defined as the level of some counterfactual costly outcome Y_0.

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 observable pre-treatment characteristics 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 reduce 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 response, 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, I 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 frequently counts or indicators for crime or harmful health and welfare outcomes. So define $\mathbf{Y}$ such that all elements $Y \geqslant 0$, and decreases in each Y are socially beneficial. Different types of people, indexed by θ, have different Y_0s and may respond differently to treatment. So in a potential outcomes framework, each row of the treatment effect vector, $\beta_{\theta,Y} = E[Y_1(\theta)] - E[Y_0(\theta)]$, may vary by person type 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.[20] 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 beneficial than 1 fewer event from a baseline of 1 (i.e., $|-3C_Y| > |-C_Y|$), even though the former is a 25 percent change and the latter is a 100 percent decline. Final conclusions about the social costs of the different behaviors SYEP affect—crime by type, health, mortality, etc.—involves specifying a social welfare function, requiring normative judgments beyond the scope of this paper. So here I 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 one counterfactual outcome, Y_0. Fig. 1 shows three stylized examples of how responsiveness to treatment could vary by risk of this outcome, plotting theoretical variation in β_θ 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_θ getting

[20] 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. I return to this point in the discussion below.

smaller when the population served is at almost no risk of the negative outcome. Here, policymakers would generate equivalent social gains 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, $\beta_\theta = -\alpha Y_0(\theta), 0 < \alpha \leqslant 1$. Here, policymakers should want to serve individuals as far to the right of the graph as possible; bigger Y_0s 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 decision-making.[21] 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 counterfactual outcome level

Estimating this relationship empirically is challenging, because Y_0 is not observed for the treatment group. I take two different approaches to understanding the relationship between Y_0 and β_θ in the SYEP setting.[22] First, I 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 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 $\widehat{Y}_0$, 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.

I note upfront that predicting a single Y_0 and estimating treatment heterogeneity by $\widehat{Y}_0$ 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 $\widehat{Y}_0$ 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.

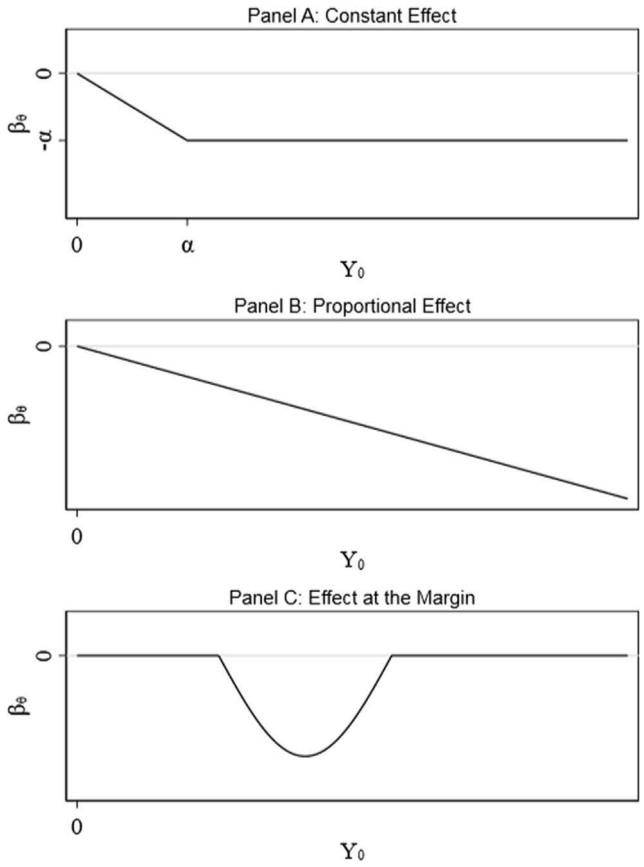

Fig. 1. Stylized Treatment Effects Relative to Counterfactual Outcomes with Different Types of Heterogeneity. *Note:* Figure shows theoretical shape of risk-responsiveness relationship under different types of treatment heterogeneity, where Y_0 is a counterfactual outcome in the absence of treatment and B_0 is the treatment effect across different types of people with different Y_0s. See Section 7.1 for discussion.

To address the latter issue, I 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, I standardize each outcome with a statistically significant main effect and generate an unweighted average of these outcomes by city.[23] In Chicago this includes drug and other arrests; in Philadelphia it includes incarceration, total arrests, and child protective services.[24]

[21] 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; Bloom, 2010; Card et al., 2017; Cave et al., 1993; Doleac et al., 2020; Heinrich et al., 2013; MDRC, 1980). Some, like the JTPA, even have adverse crime effects on those already at elevated risk of crime (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; Roder and Elliott, 2020; Schaberg and Greenberg, 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.

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

[23] In this 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 I can not include behavioral health data in this exercise. I limit the index to statistically significant main effects to avoid diluting the index with additional noise, but Appendix I.1 shows that I get a generally similar pattern of results when using an index of all the available measures of socially costly outcomes. The pattern is clearer in Philadelphia than in Chicago, but with less precision than the results in the main text.

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

Each variable is standardized on the control group, averaged together, then re-standardized so that the standard deviation of the index is 1.

Table 4 reports the overall ITT effect on the index, a 0.065 standard deviation decline in Philadelphia and 0.051 decline in Chicago, as well as the results of the endogenous stratification exercise.[25] Consistent with the pattern in the main crime results, treatment point estimates are considerably more negative for the groups with higher predicted counterfactual 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. This suggests that those who benefit the most are least likely to take-up the program on their own, and that the actual differences in responsiveness conditional on take-up are even larger than the differences in ITT point estimates would suggest.[26] 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 that may be linked to the cities' different recruiting strategies.[27]

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.[28] 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 SYEP 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, I 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 two separate experiments in this paper, plus published main and subgroup effects from two prior OSC+ cohorts, NYC, and Boston experiments. I 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 $\overline{Y}_0$.[29]

This approach 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 $\overline{Y}_0$ and β_Y is not directly interpretable, since a one-unit increase in $\overline{Y}_0$ 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 $\boldsymbol{\beta_Y}$ 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.

Fig. 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. One might worry about selecting estimates based on statistical significance; since with these outcomes, variances grow as control means rise, true effects would have to be bigger to reach statistical significance on the right side of the graph, were sample size and take-up rates equal. Appendix Section I.4 discusses this issue in detail and shows the result holds both when accounting for the variance in each estimate and when including all main effects regardless of statistical significance.

Though the control mean in the figure 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 me to show the same set of relationships across studies, and it still captures the basic relationship between the risk level of a group and its response to treatment, inclusive of take-up decisions. Note that if take-up decisions are correlated with treatment effects, the variation in the displayed LATEs may be a function of who the compliers are in each group; it will not necessarily correspond to var-

[25] I 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.

[26] This pattern raises the question of whether doing the reverse exercise—estimating treatment effects across the distribution of how likely someone is to take up the program, as in the marginal treatment effects (MTE) literature (Brinch et al., 2017; Heckman and Vytlacil, 2005; Kowalski, 2021; Mogstad et al., 2018; Walters, 2018), rather than across the distribution of $\widehat{Y}_0$—could uncover additional insights. Doing this across different studies rather than within a single study may provide particular insight into external validity (Kowalski, 2022). In practice, however, the key assumption in this literature that MTEs are linear or monotonic in the propensity to take-up is unlikely to hold in this setting; see Appendix Section I.2 for further discussion about how this literature can inform the current analysis.

[27] See the discussion in Heller and Bhanot (2021), a companion project which explores different barriers to take-up. That paper uses 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.

[28] The power issue may be part of the reason that Davis and Heller (2020) did not find statistically significant treatment heterogeneity on crime outcomes within previous studies of OSC+.

[29] Table 1 summarizes the details of the different programs, and Appendix I.3 lists which estimates are included in each panel. To ease interpretation, I focus only on the outcomes where negative effects are desirable. This includes crime, incarceration, and mortality effects in Davis and Heller (2020), Gelber et al. (2016) and Modestino (2019). The Boston paper reports only the ITT effects; I use the reported take-up rate to scale the ITT, backing out the LATE. One could also do this exercise with the positive educational effects in Leos-Urbel, 2014; Schwartz et al., 2021, and Modestino and Paulsen, 2022, but I 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 the 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.

Table 4
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, with 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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

iation in average treatment effects by group (i.e., if everyone were forced to participate). Appendix Figure A.2 shows a very similar relationship using the ITTs from each study, indicating that the pattern in Fig. 2 is not solely due to the differences in take-up rates across groups, though it may still reflect who decides to participate.

Panel A of Fig. 2 starts with the significant main effects in this paper across outcomes and cities. This focuses on the variation in $\overline{Y}_0$ 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 with larger SYEP-driven declines in 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. 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 H presents and discusses the substantive subgroup results for WorkReady and OSC+ 2015.[32] Here I focus on the overall pattern between subgroup differences in 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 (Davis and Heller, 2020; Modestino, 2019), 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 Fig. 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 prevalence 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

[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 I avoid plotting estimates that are purely linear combinations of each other (e.g., if I show effects for OSC+ 2012 and 2013 studies separately, I do not also show the pooled estimate).

[32] I am 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. Black 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 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.

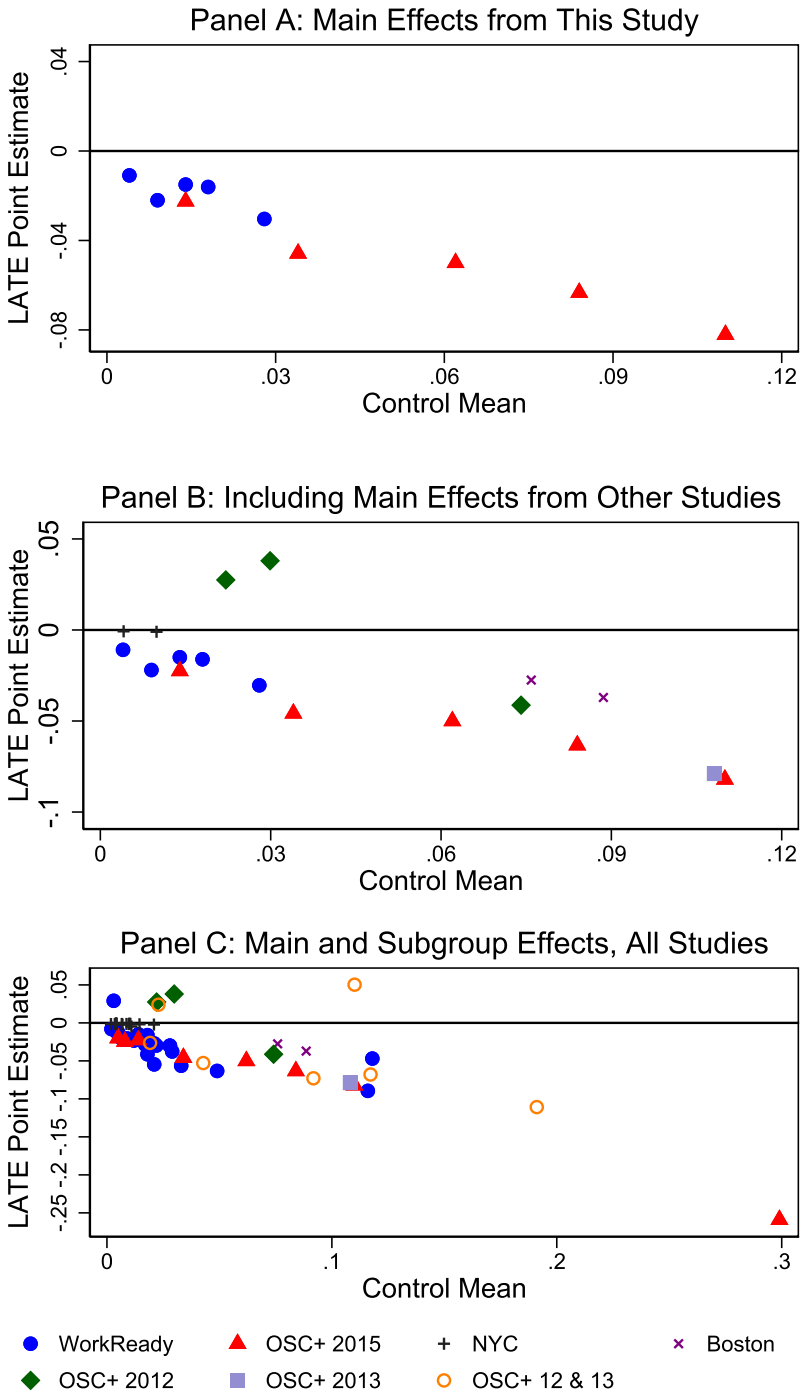

Fig. 2. Size of LATEs Relative to Control Means. *Note:* Point estimates and control means taken from this paper, Davis and Heller (2020), Gelber et al. (2016), and Modestino (2019). See text and Appendix I for details.

does mean that when SYEPs change outcomes, those changes are bigger when the outcomes are more prevalent. As discussed in Appendix I.4, the negative relationship between control means and treatment effects is robust to adjusting for each estimate's variance and to including null main effects.

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, see discussion in Appendix H). So while it is worth considering how careful SYEP targeting and program adjustments may help to minimize those increases, the overall declines in outcomes like mortality and violent crime are likely dominate any cost-benefit calculation.

7.3. Interpreting effect heterogeneity by risk level

Overall, the data seem strikingly consistent with Panel B of Fig. 1, suggesting we might extrapolate the absolute magnitude of SYEPs' 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, and both emphasize that this relationship is descriptive, not causal. Anything correlated with higher $\mathbf{Y_0}s$, including program design and implementation details that differ for higher-risk groups, could be driving the relationship.

Note it is also possible 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 Fig. 1.

Regardless of the reason, the linear relationship between risk and responsiveness holds a useful lesson for targeting SYEPs. Across all experimental studies of these programs, groups with higher counterfactual 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 SYEP participants 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 (though it may also require additional costs to serve more disconnected populations). 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 Fig. 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[33]). But given the high social costs of the type of program effects documented in this paper, 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 justice and equity concerns as well.

8. Conclusion

Variation in what treatment looks like across scales and setting—program structure, staff training or experience, counterfactual opportunities, and so forth—as well as variation in who participates 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 (see Appendix C).

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. Although treatment heterogeneity does not seem related to program structure or delivery, it is related to participant risk level. When SYEPs improve socially costly outcomes, they do so more for youth at higher risk of those outcomes. 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 in absolute magnitude, 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. It is important not to generalize too far out of sample; for example, it seems unlikely that a 6–8 week program would do much to reduce severe gun violence for those at extremely elevated risk of shooting involvement. But for the outcomes that respond to treatment among populations where SYEPs have been tested, 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

[33] 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.

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 family and health outcomes respond would be valuable. All of the major experiments on SYEPs have occurred in large cities, where the programs are quite widespread. But as Ross and Kazis (2016) point out, these results may not generalize to smaller cities or rural areas that lack the infrastructure for program administration. A better understanding of how the basic program approach changes when implemented outside large cities would help assess the potential for broader replicability.

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—at least up to the point where there are too few people at high enough risk of crime and violence to generate social benefits that outweigh program costs. For policymakers who wish to reduce social inequality efficiently, identifying other approaches that both replicate across contexts and generate the biggest benefits for the people facing the most challenges should be a priority. More broadly, the paper demonstrates how studying treatment variation and individual heterogeneity across multiple contexts has the potential to inform public spending decisions about future investments more effectively than just establishing which novel interventions "work."

Acknowledgements

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, Michael Ricks, and Basit Zafar for extremely helpful comments. I am 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. I 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+).

Appendix A. Supplementary material

Supplementary information, analysis, and results associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jpubeco.2022.104617.

References

Abadie, Alberto, 2003. Semiparametric instrumental variable estimation of treatment response models. J. Econometrics 113 (2), 231–263.
Abadie, Alberto, Chingos, Matthew M., West, Martin R., 2018. Endogenous Stratification in Randomized Experiments. Rev. Econ. Stat. 100 (4), 567–580.
Aizer, Anna, Doyle, Joseph J., 2015. Juvenile Incarceration, Human Capital and Future Crime: Evidence from Randomly-Assigned Judges. Q. J. Econ., 759–804.
Allcott, Hunt, 2015. Site selection bias in program evaluation. Q. J. Econ. 130 (3), 1117–1165.
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. J. Am. Stat. Assoc. 103 (484), 1481–1495.
Athey, Susan, Imbens, Guido, 2017. The Econometrics of Randomized Experiments. In: Banerjee, Abhijit Binayak, Duflo, Esther (Eds.), Handbook of Field Experiments. North-Holland.
Benjamini, Yoav, Hochberg, Yosef, 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy. Stat. Soc. Ser. B (Methodological) 57 (1), 289–300.
Berk, R.A., Lenihan, K.J., Rossi, P.H., 1980. Crime and Poverty - Some Experimental Evidence from Ex-Offenders. Am. Sociol. Rev. 45 (5), 766–786.
Bhatt, Monica P., Guryan, Jonathan, Ludwig, Jens, Shah, Anuj K., 2021. Scope Challenges to Social Impact. NBER Working Paper No. 28406..
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, Bos, Johannes M., 1997. The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study. J. Hum. Resour. 32 (3), 549–576.
Brinch, Christian N., Mogstad, Magne, Wiswall, Matthew, 2017. Beyond LATE with a discrete instrument. J. Polit. Econ. 125 (4), 985–1039.
Card, David, Kluve, Jochen, Weber, Andrea, 2017. What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations. J. Eur. Econ. Assoc. 16 (3), 894–931.
Cave, George, Bos, Hans, Doolittle, Fred, Toussaint, Cyril, 1993. JOBSTART: Final Report on a Program for School Dropouts. MDRC..
Charles, Kerwin Kofi, Luoh, Ming Ching, 2010. Male Incarceration, the Marriage Market, and Female Outcomes. Rev. Econ. Stat. 92 (3), 614–627.
Davis, Jonathan M.V., Guryan, Jonathan, Hallberg, Kelly, Ludwig, Jens, 2017. The Economics of Scale-Up. NBER Working Paper No. 23925..
Davis, Jonathan M.V., Heller, Sara B., 2020. Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs. Rev. Econ. Stat. 102 (4), 664–677.
Dobbie, Will, Goldin, Jacob, Yang, Crystal S., 2018. The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges. Am. Econ. Rev. 108 (2), 201–240.
Doleac, Jennifer L., Temple, Chelsea, Pritchard, David, Roberts, Adam, 2020. Which prisoner reentry programs work? Replicating and extending analyses of three RCTs. Int. Rev. Law Econ. 62..
Fein, David, 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, Kessler, Judd, 2016. The Effects of Youth Employment: Evidence from New York City Lotteries. Quart. J. Econ. 131, 423–460.
Gleicher, Lily, 2017. Juvenile justice in Illinois, 2015. Illinois Criminal Justice Information Authority..
Goncalves, Felipe, Mello, Steven, 2021. A Few Bad Apples? Racial Bias in Policing. Am. Econ. Rev. 111 (5), 1406–1441.
Heckman, James J., Smith, Jeffrey, Clements, Nancy, 1997. Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. Rev. Econ. Stud. 64 (4), 487–535.
Heckman, James J., Vytlacil, Edward, 2005. Structural equations, treatment effects, and econometric policy evaluation 1. Econometrica 73 (3), 669–738.
Heinrich, Carolyn J., Mueser, Peter R., Troske, Kenneth R., Jeon, Kyung-Seong, Kahvecioglu, Daver C., 2013. Do Public Employment and Training Programs Work?. IZA J. Lab. Econ. 2..
Heller, Sara B., 2014. Summer jobs reduce violence among disadvantaged youth. Science 346, 1219–1223.
Heller, Sara B., Bhanot, Syon, 2021. Overcoming Application and Take-Up Barriers for Summer Youth Employment Programs..

Hinton, Elizabeth, Henderson, LaShae, 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, Stoll, Michael A., 2006. Perceived Criminality, Criminal Background Checks, and the Racial Hiring Practices of Employers. J. Law Econ. 49 (2), 541–580.
Jepsen, Christopher, Rivkin, Steven, 2009. Class Size Reduction and Student Achievement: The Potential Tradeo? Between Teacher Quality and Class Size. J. Hum. Resour. 44 (1), 223–250.
Kessler, Judd B., Tahamont, Sarah, Gelber, Alexander M., 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., Katz, Lawrence F., 2007. Experimental Analysis of Neighborhood Effects. Econometrica 75, 83–119.
Kowalski, Amanda E., 2021. Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform. The Review of Economics and Statistics (forthcoming).
Kowalski, Amanda E., 2022. Behavior within a clinical trial and implications for mammography guidelines. Review of Economic Studies (forthcoming).
Leos-Urbel, Jacob, 2014. What is a Summer Job Worth? The Impact of Summer Youth Employment on Academic Outcomes. J. Policy Analysis Manage. 33, 891–911.
MDRC, 1980. Summary and Findings of the National Supported Work Demonstration..
Modestino, Alicia Sasser, 2019. How Do Summer Youth Employment Programs Improve Criminal Justice Outcomes, and for Whom? J. Public Policy Anal. Manage..
Mogstad, Magne, Santos, Andres, Torgovitsky, Alexander, 2018. Using instrumental variables for inference about policy relevant treatment parameters. Econometrica 86 (5), 1589–1619.
Modestino, Alicia Sasser, Paulsen, Richard, 2022. School's Out: How Summer Youth Employment Programs Impact Academic Outcomes. Education Finance and Policy (forthcoming)..
Mueller-Smith, Michael, 2015. The Criminal and Labor Market Impacts of Incarceration. University of Michigan Working Paper..
Ridgeway, Greg, MacDonald, John, 2009. Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops. J. Am. Stat. Assoc. 104 (486), 661–668.
Roder, Anne, Elliott, Mark, 2020. Stepping Up: Interim Findings on JVS Boston's English for Advancement Show Large Earnings Gains. Economic Mobility Corporation..
Ross, Martha, Kazis, Richard, 2016. Youth Summer Jobs Programs: Aligning Ends and Means. The Brookings Institution.
Schaberg, Kelsey, Greenberg, David H., 2020. Long-Term Effects of a Sectoral Advancement Strategy: Costs, Benefits, and Impacts from the Work Advance Demonstration. MDRC..
Al-Ubaydli, Omar, List, John A., Suskind, Dana, 2020. The Science of Using Science: Towards an Understanding of the Threats to Scaling Experiments. International Economic Review 61 (4)..
Schwartz, Amy Ellen, Leos-Urbel, Jacob, Wiswall, Matthew, 2021. Making Summer Matter: The Impact of Youth Employment on Academic Performance. Quantitative Economics 12 (2), 477-504..
Walters, Christopher R., 2018. The demand for effective charter schools. J. Polit. Econ. 126 (6), 2179–2223.
Weiss, Michael J., Bloom, Howard S., Verbitsky-Savitz, Natalya, Gupta, Himani, Vigil, Alma E., Cullinan, Daniel N., 2017. How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. J. Res. Educ. Effectiv. 10 (4), 843–876.
Westfall, Peter H., Young, S. Stanley, 1993. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley-Interscience..