NBER WORKING PAPER SERIES

PREVENTING YOUTH VIOLENCE AND DROPOUT: A RANDOMIZED FIELD EXPERIMENT

Sara Heller
Harold A. Pollack
Roseanna Ander
Jens Ludwig

Working Paper 19014
http://www.nber.org/papers/w19014

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2013

This project was supported by the University of Chicago's Office of the Provost, Center for Health Administration Studies, and the School of Social Service Administration, as well as NICHD award R21HD061757, CDC grant 5U01CE001949-02 to the University of Chicago Center for Youth Violence Prevention, grants from the Joyce, MacArthur, McCormick, Polk, and Spencer foundations, the Exelon corporation, and the Chicago Community Trust, and visiting scholar awards to Jens Ludwig from the Russell Sage Foundation and LIEPP at Sciences Po. We are grateful to the staff of Youth Guidance and World Sport Chicago (the two non-profit organizations that implemented the intervention we study here), to Wendy Fine of Youth Guidance, who designed and implemented required program data systems, to the Chicago Public Schools, to the Illinois Criminal Justice Information Authority for providing Illinois Criminal History Record Information (CHRI) data through an agreement with the Illinois State Police, and to Ellen Alberding, Jon Baron, Dan Black, Laura Brinkman, Carol Brown, Kerwin Charles, Philip Cook, Stephen Coussens, Hon. Richard M. Daley, Christine Devitt Westley, Ken Dodge, Steve Gilmore, Jonathan Guryan, Hon. Curtis Heaston, Ron Huberman, Brian Jacob, Rachel Johnston, Ilyana Kuziemko, Ben Lahey, Ann Marie Lipinski, John MacDonald, Sonya Malunda, Jeanne Marsh, Michael Masters, Michael McCloskey, Al McNally, Ernst Melchior, Douglas Miller, Michelle Morrison, Duff Morton, Sendhil Mullainathan, Mark Myrent, Derek Neal, Stacy Norris, Amy Nowell, Devah Pager, Steve Raudenbush, Sean Reardon, Thomas Rosenbaum, Anuj Shah, David Showalter, Sebastian Sotelo, Laurence Steinberg, Ashley Van Ness, Nina Vinik, Paula Wolff, and Sabrina Yusuf for valuable assistance and suggestions. We also thank seminar participants at the Boeing Corporation, Case Western University, Columbia University, Duke University, Erasmus University, Harvard University, the MacArthur Foundation, National Bureau of Economic Research, New York City Department of Probation,

Preventing Youth Violence and Dropout: A Randomized Field Experiment
Sara Heller, Harold A. Pollack, Roseanna Ander, and Jens Ludwig
NBER Working Paper No. 19014
May 2013
JEL No. I24,I3,K42

ABSTRACT

Improving the long-term life outcomes of disadvantaged youth remains a top policy priority in the United States, although identifying successful interventions for adolescents – particularly males – has proven challenging. This paper reports results from a large randomized controlled trial of an intervention for disadvantaged male youth grades 7-10 from high-crime Chicago neighborhoods. The intervention was delivered by two local non-profits and included regular interactions with a pro-social adult, after-school programming, and – perhaps the most novel ingredient – in-school programming designed to reduce common judgment and decision-making problems related to automatic behavior and biased beliefs, or what psychologists call cognitive behavioral therapy (CBT). We randomly assigned 2,740 youth to programming or to a control group; about half those offered programming participated, with the average participant attending 13 sessions. Program participation reduced violent-crime arrests during the program year by 8.1 per 100 youth (a 44 percent reduction). It also generated sustained gains in schooling outcomes equal to 0.14 standard deviations during the program year and 0.19 standard deviations during the follow-up year, which we estimate could lead to higher graduation rates of 3-10 percentage points (7-22 percent). Depending on how one monetizes the social costs of crime, the benefit-cost ratio may be as high as 30:1 from reductions in criminal activity alone.

Sara Heller
University of Chicago
Harris School of Public Policy
1155 East 60th Street
Chicago, IL 60637
sbheller@uchicago.edu

Harold A. Pollack
University of Chicago
School of Social Service Administration
969 East 60th Street
Chicago, IL 60637
haroldp@uchicago.edu

Roseanna Ander
University of Chicago Crime Lab
720 North Franklin Street, Suite 400
Chicago, IL 60654
rander@uchicago.edu

Jens Ludwig
University of Chicago
1155 East 60th Street
Chicago, IL 60637
and NBER
jludwig@uchicago.edu

I. INTRODUCTION

Improving the long-term life outcomes of disadvantaged youth remains a top policy priority in the United States. The average four-year high school graduation rate in the 50 largest urban school districts in America is just 53 percent (Swanson 2009). Nearly 70 percent of black male dropouts will spend time in prison by their mid-30s (Western & Pettit 2010). Among males ages 15-24, the homicide 2010 rate for blacks was 18 times that of whites (75 vs. 4/100,000). Because homicide victims are disproportionately young, more years of potential life are lost to homicide among black males than to the nation's leading overall killer – heart disease.[1]

Long-term progress in addressing these problems has been limited, in part because finding ways to improve outcomes for disadvantaged youth (particularly males) has proven challenging.[2] Despite technological changes that have increased the demand for educated workers over time (Goldin & Katz 2008), the high school graduation rate in America has not changed much since the 1970s.[3] While mortality rates from almost every major leading cause of death have declined dramatically over the past half century, the homicide rate today is not much different from what it was in 1950 – or even in 1900 (Pinker 2011, p. 92).[4]

The persistence of these problems, and the limited success of previous social-policy efforts, often lead to the conclusion that very intensive intervention is needed to overcome the

[1] Figures are for years of potential life by 65. http://www.cdc.gov/injury/wisqars/fatal_injury_reports.html
[2] For example, the U.S. Department of Education's What Works Clearinghouse (WWC) does not give a single dropout-prevention program its top rating of "strong effects" (defined as several randomized experiments or quasi-experiments all pointing in the same direction, or one large randomized experiment). The Coalition for Evidence-Based Policy does not list a single program for addressing high school graduation rates among its "Top Tier" of programs. The evidence about how to reduce youth violence is not much stronger; see Appendix A for details.
[3] While Heckman and LaFontaine (2010) show high school graduation rates were flat in the U.S. as a whole from the 1970s through 2000, Murnane (2013) shows graduation rates have increased over the past 10 years. But the share of those born in 1986-90 with a diploma is 84% - not much higher than the 81% for those born in 1946-50.
[4] From 1950 to 2005, mortality rates declined by 45 percent from all causes, 64 percent from deaths due to heart disease, 74 percent from cerebrovascular diseases, 58 percent from influenza and pneumonia, 50 percent for unintentional injuries, and 20 percent for chronic liver disease and cirrhosis. Aside from homicide, the two other exceptions to the long-run decline are cancer and diabetes mellitus (National Center for Health Statistics 2009).

powerful "root causes" that may drive adverse youth outcomes (see, for example, Garbarino 1999, Chapter 7). The social conditions most often implicated in discussions about adolescent outcomes – economic disadvantage, under-performing public schools, the way children are parented and socialized growing up – are all difficult to change, and their consequences may be challenging to overcome. A related worry is that the effects of adverse social conditions may be too entrenched by adolescence, so that interventions should focus more on early childhood when people may be more malleable (Knudsen, et al. 2006; Shonkoff & Phillips 2000).

However, there may be a different piece of this problem that lends itself to lower-cost policy intervention: The effects of disadvantaged social conditions on youth outcomes may be at least partly mediated by errors in judgment and decision-making. Kahneman (2011) notes that all of us rely on automatic, intuitive decision-making, which is sometimes generated from mistaken inferences and beliefs. However the consequences may vary greatly by circumstance. The likelihood of holding specific biased beliefs may also vary systematically within the population. For example, Dodge and Pettit (1990) show that one cause of aggressive behavior is hyper-vigilance to threat cues and the tendency to over-attribute malevolent intent to others, or "hostile attribution bias." Like other theories of crime, social conditions play a role: This bias seems to be more common among those from disadvantaged backgrounds, due partly to elevated risk of having experienced abuse growing up. But now there is a mechanism that might be addressed directly, not just a root cause. Many schooling decisions may stem from similar errors, given that dropout is often precipitated by a disciplinary action or conflict with a teacher.

Our paper presents the results of a large-scale randomized controlled trial (RCT) that took place in 18 Chicago Public Schools (CPS) in some of the city's most disadvantaged and dangerous south and west side neighborhoods. We randomly assigned 2,740 male youth grades

7-10 to program or control conditions for the 2009-10 academic year. The intervention, called "Becoming a Man" (BAM), was run by two local nonprofits, Youth Guidance (YG) and World Sport Chicago (WSC). About half the youth assigned to treatment participated; the average participant attended 13 one-to-two hour sessions.

The intervention's components include regular exposure to pro-social adults, a key ingredient for almost any social-policy intervention, after-school programming, and – perhaps the most novel ingredient – cognitive behavioral therapy (CBT). CBT is a short-duration intervention from psychology that helps people recognize and reduce unhelpful automatic behaviors and biased beliefs – to promote "thinking about thinking" (meta-cognition). Since the 1970s, CBT has been used to address mental health disorders such as substance abuse, anxiety, and depression, and indeed can be more effective than anti-depressant drug treatment (Rush, et al. 1977). Since then, there has been growing practitioner interest in using CBT to address socially-costly behaviors, though little good evidence currently exists about effects on those behaviors of greatest policy concern such as delinquency, violence, and dropout.[5] We measure those outcomes during the program and a follow-up year using administrative data, which are not subject to the same sample attrition and misreporting problems that often afflict survey data.

Using random assignment as an instrument for participation, we find that participation reduced violent crime arrests by 8.1 arrests per 100 youth over the course of the program year, a decline of 44 percent relative to participants' control group counterparts. Arrests in our "other" (non-violent, non-property, non-drug) category decreased by 11.5 arrests per 100 youth during

[5] For example, a meta-analysis by Drake, Aos and Miller (2009) identified just a single "high-quality" experiment carried out with youth, by Armstrong (2003), which found no significant effects on recidivism rates among juveniles in a Maryland detention center. The lack of detectable impacts could mean that the intervention "doesn't work," but could also be due instead to the modest sample size (110 treatment youth and 102 controls), or to the fact that the treatment and control groups do not in fact appear to be comparable with respect to key baseline characteristics such as share African-American, equal to 67 and 48 percent, respectively (see Appendix A for a review of this literature).

the program year, a decline of 36 percent, due mostly to reductions in weapons offenses together with vandalism and trespassing. While these large arrest impacts did not persist, participation also led to lasting gains in an index of schooling outcomes equal to 0.14 standard deviations (sd) in the program year and 0.19sd in the follow-up year. Our sample is too young to have graduated, but based on correlations from previous longitudinal studies of CPS students, we estimate our schooling impacts could imply gains in graduation rates of 3-10 percentage points (7-22 percent). With a cost of $1,100 per participant, depending on how we monetize the social costs of violent crime, the benefit-cost ratio is up to 30:1 just from effects on crime alone.

The size of these effects, together with the modest "dosage," suggests that even serious youth outcomes may be more elastic to policy intervention than previous research would suggest. While our reliance on administrative data necessarily limits our ability to isolate mechanisms, the fact that previous programs that provide interactions with pro-social adults or after-school activities tend not to show similarly large effects is at least suggestive evidence that the novel ingredient here – CBT – may be important. Our results are not due just to "incapacitation" of youth after school, since arrest impacts are at least as large on days when after-school programming is not offered. We also have access to CPS student surveys that suffer from low response rates, but provide at least suggestive evidence that the intervention may have improved measures of perseverance ("grit") and items related to conflict resolution and peer relationships.

As one juvenile detention staff member told us: "20 percent of our residents are criminals, they just need to be locked up. But the other 80 percent, I always tell them – if I could give them back just ten minutes of their lives, most of them wouldn't be here."[6] Our results suggest that it is possible to generate sizable changes in outcomes by helping disadvantaged

[6] Personal communication, Darrien McKinney to Jens Ludwig, Sendhil Mullainathan, and Anuj Shah, 10/18/2012.

youth recognize their own thinking patterns and make better decisions during those crucial ten-minute windows.

The next section briefly reviews some key characteristics of youth violence and dropout behavior as a way to highlight the potential pathways through which our intervention may affect youth outcomes. Section three describes the intervention we study. We discuss our study sample in section four, program participation and "cross-over" in section five, data and outcomes in section six, and analytic methods in section seven. Our main findings for crime and schooling are in sections eight and nine. Extensions are in section ten, including evidence of robustness to how we handle missing data and multiple comparisons, results by treatment arm, and tests for treatment heterogeneity across students and schools. Section eleven discusses the evidence we can assemble about mechanisms; section twelve presents benefit-cost estimates; and the final section discusses limitations and implications.

II. YOUTH VIOLENCE AND DROPOUT

The factors that contribute to adverse youth outcomes, and the pathways through which the intervention we study might help, are easiest to see with a concrete example. While examples from education are plentiful, youth violence illustrates the key points in a particularly sharp way.

At 3pm on Saturday, June 2, 2012, in the South Shore neighborhood just a few miles from the University of Chicago, two groups of teens were arguing in the street about a stolen bicycle. As the groups began to separate, someone pulled out a handgun and fired, hitting a 16-year-old named Jamal Lockett in the chest. Lockett was rushed up Lake Shore Drive to

Northwestern Hospital where he was pronounced dead. Two weeks later, prosecutors filed first-degree murder charges against the alleged shooter, Kalvin Carter – 17 years old.[7]

The example illustrates many of the familiar social conditions thought to contribute to youth violence: Chicago's violence is disproportionately concentrated in economically and racially segregated areas like South Shore, where 95 percent of residents are African-American, 27 percent are poor, and a majority of households with children contain only one parent. Violence in general is disproportionately committed by young people when they are not under adult supervision – particularly weekends and the afternoon hours when school lets out.[8]

The example is also representative with respect to its motivation, which highlights the potential impact of CBT interventions that reduce errors in judgment and decision-making. While media portrayals emphasize strategic, instrumental violence (for example, the shootings committed by Snoop Pearson and Chris Partlow as part of Marlo Stanfield's drug war against Avon Barksdale in *The Wire*), as suggested by our example, this is not true of most violent events: In Chicago, the site of our study, police believe that roughly 70 percent of homicides stem from "altercations," compared to only about 10 percent from drug-related gang conflicts.[9]

It is possible that many altercations such as the one described above escalate into violence because youth are making intuitive, even automatic decisions, which Kahneman (2011) suggests are common – but may not be adaptive in all circumstances, such as when a gun is readily at hand. At 3pm on June 2 on the south side of Chicago, is Kalvin Carter thinking about 3:01 – or even consciously thinking at all, for that matter? Automatic, intuitive decision-making

---

[7] See http://chicago.cbslocal.com/2012/06/03/dispute-over-bicycle-blamed-in-teens-fatal-shooting/ and http://articles.chicagotribune.com/2012-06-15/news/chi-kalvin-carter-17-charged-with-killing-jamal-lockett-16-20120615_1_teens-shot-riverdale-weapon-charges

[8] *OJJDP Statistical Briefing Book*. Online. http://www.ojjdp.gov/ojstatbb/offenders/qa03301.asp?qaDate=2008. Released on December 21, 2010 ; accessed February 14, 2013.

[9] In 2011, there were 433 Chicago homicides total. The motivations in 121 cases were unknown to the police; 219 of the remaining 312 homicides were attributed by the police to an altercation (CPD 2011b).

is also susceptible to systematic biases, partly because the brain's automatic "system" tends to emphasize explanations that are coherent rather than necessarily correct. Examples of such errors include hostile attribution bias (Kalvin Carter may have taken the denial of knowledge about the stolen bicycle as evidence of deceit or disrespect, not innocence), confirmation bias (focusing on information that confirms one's preconceptions – perhaps ignoring conciliatory words by all but one member of the other group), or catastrophizing (the tendency to think negative events are even more negative than they are – perhaps Kalvin Carter thought "literally nothing is worse than letting down my friends").

It is also not hard to see how overly automatic behavior and biased beliefs could lead to trouble in school. Hostile attribution bias cannot be helpful in a world in which teachers sometimes raise their voices at students to start class or maintain discipline, and might contribute to the sorts of disciplinary actions that often lead to school disengagement and eventually dropout.[10] Catastrophizing increases the risk that a student who has just received a low grade on an exam might conclude he is incapable of high-school-level academic work, and give up.

The possible role of judgment and decision-making errors in explaining adverse youth outcomes does not rule out a role for deficits in "non-cognitive" or "social-cognitive" skills as well.[11] But if the field's experiences trying to improve academic skills are any guide, remediating social-cognitive skill deficits may turn out to require very intensive intervention. Addressing

[10] A qualitative study of youth in the Moving to Opportunity (MTO) mobility experiment finds that disengagement from school often follows a disciplinary action – that is, one mistake or over-reaction interacting with a peer or teacher can start a process that ends in dropout (Clampet-Lundquist, DeLuca & Edin 2012). Rumberger (2001) reviews data from the National Education Longitudinal Study of 1988 eighth graders and finds 39 percent said they dropped out because they were 'failing school;' 29 percent 'could not get along with teachers.' Another 49 percent said they 'did not like school,' and 27 percent of dropouts cited getting a job as a reason.

[11] Research in psychology and economics shows non-academic skills are correlated with a range of life outcomes. (see, for example, Borghans, et al. 2007; Bowles, Gintis & Osborne 2001; Cunha & Heckman 2007; Dodge 2003; Heckman & Rubinstein 2001; Heckman, Stixrud & Urzua 2006; Moffitt, et al. 2011; Monahan, et al. 2009).

judgment and decision-making errors could potentially be amenable to lighter-touch (and hence less costly) intervention, since the goal is largely recognition and awareness – epiphanies.

III. INTERVENTION

The intervention we study here, called "Becoming a Man" (BAM), was developed and implemented by two Chicago-area non-profit organizations, Youth Guidance (YG) and World Sport Chicago (WSC). It includes in-school and after-school programming that expose youth to pro-social adults, occupy them during the high-risk hours after school, and implement aspects of what psychologists call cognitive behavioral therapy (CBT) designed to get youth to "think about thinking" (promote meta-cognition): that is, to recognize situations in which automatic, intuitive decision-making may lead to trouble and to recognize (and correct) biased beliefs or interpretations of their experiences (Beck 2011).

The in-school treatment offered the chance to participate in up to 27 one-hour, once-per-week group sessions during the school day over the school year. The intervention is delivered in groups to help control costs, with groups kept small (assigned groups of no more than 15 youth and a realized average youth-to-adult ratio of 8:1) to help develop relationships. Students skip an academic class in order to participate in the program, which is one of the draws for many youth to attend. The program is manualized and can be delivered by college-educated people without specialized training in psychology or social work, although YG had a preference for such training in selecting program providers. From observing sessions, it also seems clear that another skill essential to success is the ability to keep youth engaged.

The curriculum includes standard elements of CBT (Beck 2011), such as a common structure to most sessions that starts with a self-analysis ("check in") to help identify problematic thoughts or behaviors to be addressed. Participants discuss a cognitive model emphasizing that

emotional reactions to events are endogenous and often influenced by automatic thoughts, and are taught relaxation techniques to help avoid overly automatic reactions ("out of control" behavior). Stories, movies, and metaphors are used to illustrate unhelpful automatic behaviors and biased beliefs at work in the lives of others. Youth are taught to use "behavioral experiments" to empirically test their biased beliefs, both during program sessions and as homework in between sessions, with a special emphasis on common social-information-processing errors and problems around perspective-taking, such as catastrophizing and a focus on overly narrow, short-term goals. Because monitoring automatic thoughts requires effort, CBT helps focus this effort by helping people recognize indicators that some maladaptive automatic thought or biased belief is being triggered. A shift to some aversive emotion is one common cue (Beck 2011). Given the common risks for this population, a key focus was on anger as a cue.

The nature of the intervention is best illustrated by example. The very first activity for youth in the program is the "Fist Exercise." Students are divided into pairs; one student is told he has 30 seconds to get his partner to open his fist. Then the exercise is reversed. Almost all youth attempt to use physical force to compel their partners to open their fists. During debrief, the group leader asks youth to explain what they tried and how it worked, pointedly noting that (as is usually the case) almost no one has *asked* their partner to open their fist. When youth are asked why, they usually provide responses such as: "he wouldn't have done it," or "he would have thought I was a punk." The group leader will then follow-up by asking: "How do you know?" The exercise is an experiential way to teach youth about hostile attribution bias. The example also shows how the program is engaging to youth who might not normally sign up for pro-social activities, because it is slightly subversive – to participate they get out of an academic class, and then the first activity winds up involving sometimes-rowdy horseplay.

The broader intervention also includes after-school programming delivered by WSC designed to both enhance program participation rates and provide youth with more opportunities to reflect on their automatic responses and decision-making. The WSC coaches all receive some training in the BAM program. WSC sessions, one-to-two hours each, include non-traditional sports (archery, boxing, wrestling, weightlifting, handball, and martial arts) that require focus, self-control, and proper channeling of aggression, and also provide youth with additional opportunities for reflection on their automatic behavior ("so after you got hit in the face during that boxing match, what were you thinking that led you to drop your hands and charge blindly?")

IV. STUDY SAMPLE AND RANDOMIZATION

Our study setting – the Chicago Public Schools (CPS) – is similar to those of many other large, urban school districts that serve disproportionately disadvantaged populations. Of the 409,000 students in the CPS system, 86 percent are low-income, and over 90 percent are racial or ethnic minorities.[12] Of students who begin 9th grade in CPS, only 51 percent graduate high school within four years, about average for large urban school systems (Swanson, 2009), and only eight percent graduate from a four-year college (Allensworth 2006). Chicago's homicide rate is well above the national average but middle of the pack for large U.S. cities.

During the summer of 2009, our team recruited 18 elementary and high schools in the Chicago Public School (CPS) system located on Chicago's low-income, racially segregated South and West sides, where the city's violent crime is disproportionately concentrated. Our study sample is essentially the 2,740 highest-risk male students in grades 7-10 in the 18 CPS study schools, after excluding students who rarely attend school (and so would not benefit from a school-based intervention) or who have serious disabilities. This sample represents around 75 percent of all male youth in grades 7-10 in the study schools (see Appendix B for details).

[12] http://www.cps.edu/About_CPS/At-a-glance/Pages/Stats_and_facts.aspx

Youth were randomized to treatment (in-school, after-school, or both) or control groups within each school, so our design is a block-randomized experiment with schools as blocks.[13] Our analyses control for school fixed-effects given the block-randomized design. One key question for our study is whether any programming affects outcomes, and secondarily, which elements matter most. It turns out that we do not have statistical power to distinguish among the three treatment arms, a problem compounded by treatment cross-over (described below). For these reasons, and because all three arms share the same larger goals, our main results focus on the effect of all three arms pooled together. We also present results by treatment arm below.

Table I shows that ours is a very disadvantaged sample. The average age at baseline was 15, with over half being old for grade. Reflecting the composition of their neighborhoods, all of our study youth are minorities – around 70 percent black, the remainder Hispanic. During the pre-randomization year (AY 2008-9), the average youth had a GPA of 1.7 on a 4.0 scale and attended 130 out of a total possible 170 school days. Over one-third of study youth had been arrested at least once prior to randomization.

The similarity of baseline characteristics between treatment and control groups in Table I suggests that randomization was successful. Of 19 total baseline covariates we examined,[14] none of the treatment-control differences is significantly different at the 5 percent level. An F-test for the joint significance of all available baseline characteristics shows we cannot reject the null hypothesis that treatment and control groups are equivalent ($F(19,2542)=1.05$, $p=.397$).

V. PROGRAM PARTICIPATION AND CROSS-OVER RATES

[13] Three of our 18 schools could not accommodate after-school programming because of logistical or space reasons, so they included only in-school and control conditions. Eight schools offered both in- and after-school treatment arms in some combination, but did not offer all three treatment arms in addition to the control group.

[14] The other baseline variables not shown in Table I are: number of in-school suspensions, number of out-of-school suspensions, and each of the number of grades earned (A through F).

Table II shows that around half of youth offered the chance to participate in program activities chose to participate. This take-up rate is consistent with other large scale social experiments (Bloom, et al. 1997; Kling, Liebman & Katz 2007) despite the fact that we randomized first (using administrative data) and then tried to consent people for program participation, rather than consenting and then randomizing, as is more common.[15] We suspect participation rates for the after-school programming are under-stated because of inadequate record keeping; below, we bound the impact of this under-reporting on our estimated effect of participating. Among participants, the average number of sessions attended is around 13.

Figure I highlights one hard-to-avoid byproduct of running social experiments in challenging circumstances like those in the CPS system: namely, some control group youth wind up participating in program activities ("control-group cross-over"). We analyze the data using original treatment or control assignments to preserve the strength of the experimental design. Figure I also shows that there is treatment-group cross-over as well. For example, fully one-third of participants among the youth assigned to receive *only* after-school programming wound up receiving in-school programming. Among those youth offered both activities, more received just in-school programming only than received both activities.

VI. OUTCOME MEASURES

Our main schooling outcomes come from longitudinal student-level CPS records for the pre-program year (AY 2008-9), program year (AY 2009-10), and follow-up year (AY 2010-11). Our sample was drawn from CPS data on the pre-program year and then matched to data from subsequent years using CPS student ID numbers. We used the post-randomization data to form a summary index of schooling outcomes, which reduces the number of hypothesis tests and so reduces risk of "false positives" (Anderson 2008; Kling, Liebman & Katz 2007; Westfall &

[15] Consent was for program participation only; outcome data is available for all youth who were randomized.

Young 1993). It also improves the statistical power available to detect effects for outcomes within a given family expected to move in a similar direction. Below we show this is indeed true of the elements of our index, which is an (unweighted) average of days present, GPA, and persistence in school (enrollment status at the end of the academic year), each normalized to Z-scores using the control group's distribution (Appendix C has additional details).[16]

While we can determine enrollment status for each student in every year, the share of students missing at least one of the other elements of our index grows over time from 10 percent in the program year to 33 percent during the follow-up year. For individuals missing some element(s) of the composite, but who have valid information for at least one component, we follow Kling, Liebman, and Katz (2007) and assign the group (treatment or control) mean for the missing elements. This approach has the advantage of using all available information and has a straightforward interpretation: it is equivalent to estimating the treatment effect on each component of the index (in standardized form) using only observations with non-missing observations, and then averaging the component-specific estimates. Below we demonstrate that our results are generally robust to alternative ways of handling missing data.

To measure criminal behavior by program participants, we use electronic arrest records (or "rap sheets") from the Illinois State Police (ISP), which were matched to our study sample for research purposes by the Illinois Criminal Justice Information Authority (ICJIA) using probabilistic matching on name and date of birth. Arrest records avoid the problem of under-reporting of criminal involvement in survey data (Kling, Ludwig & Katz 2005) but require the assumption that the intervention itself does not affect the likelihood that criminal behavior results

[16] Our index of schooling outcomes does not include standardized test scores. By design, CPS does not administer standardized tests to all grades (particularly older grades). Thus, more than half of all students in our study sample are missing test scores (similar shares for treatment and control groups). Our index also does not include administrative records on school disciplinary actions; we have received conflicting accounts from different sources about whether disciplinary actions are inconsistently reported and recorded in CPS data, even within schools.

in arrest. The ISP records capture arrests in the state going back to 1990 and include arrests of people below the age of majority within the criminal justice system (juvenile arrests), as well as to those who are above the age of majority. Local police departments are required by law to report all juvenile felony arrests to the ISP, and optionally class A and B misdemeanors. Because intervention impacts often vary greatly by crime type (Deming 2011; Evans & Owens 2007; Kling, Ludwig & Katz 2005; Lochner & Moretti 2004; Weiner, Lutz & Ludwig 2009), as do social costs, we examine arrests separately for four different offense categories: violent, property, drug, and "other" (excluding motor vehicle violations).

VII. ANALYSIS APPROACH

The main challenge in identifying the effects of social-policy interventions is selection – those youth who select into (or are selected for) participation may be systematically different from non-participants in ways that also directly affect outcomes. We overcome this problem with a randomized control group, which lets us identify the average outcomes the treatment group would have experienced had they not been offered the program.

Let Y_{ist} denote some post-program outcome for individual *i* at school *s* during post-randomization period *t*, which is a function of treatment group assignment (Z_{is}) and observed variables from ISP and CPS records measured at or before baseline ($X_{is(t-1)}$) as in equation (1) below. We control for the blocking variable with school fixed effects (γ_s). The "Intent-To-Treat effect" (ITT) captures the effect of being offered the chance to participate in the program, and is given by the estimate of π_1 in equation (1). We condition on baseline characteristics to improve precision by accounting for residual variation in the outcomes (results without baseline

covariates are similar; available upon request).[17] Our main tables present results for the treatment arms pooled together (see Section V), but later we also show results separately as well.[18]

(1) $Y_{ist} = Z_{is}\pi_1 + X_{is(t-1)}\beta_1 + \gamma_s + \varepsilon_{ist1}$

The advantage of the ITT estimand is that it fully exploits the strength of the randomized experimental design. But since not all youth participate, the ITT will understate the effects of actually participating in the program. We therefore also report the effect of participating in the program for those who actually participate, which we estimate using two-stage least squares with random assignment (Z_{is}) as an instrumental variable (IV) for participation (P_{ist}), as in equations (2) and (3) (Angrist, Imbens & Rubin 1996; Bloom 1984). This assumes treatment-group assignment has no effect on the behavior of youth who do not participate in the intervention.

(2) $P_{ist} = Z_{is}\pi_1 + X_{is(t-1)}\beta_1 + \gamma_s + \varepsilon_{ist2}$

(3) $Y_{ist} = P_{ist}\pi_2 + X_{is(t-1)}\beta_2 + \gamma_s + \varepsilon_{ist3}$

If youth respond differently to the intervention, then because some of our controls wind up in the program, π_2 is a local average treatment effect (LATE) – the effect of the program (treatment) on those whose treatment receipt is affected by being assigned to the treatment rather than control group, or "compliers" (Angrist, Imbens & Rubin 1996; Imbens & Rubin 1997; Imbens & Angrist 1994)). Because control cross-over rates are low, we expect the LATE should be fairly close to the effect of treatment on the treated. The IV results are essentially equal to the

[17] Specifically, we control for the following variables from the 2008-9 academic year: total days present; number of in- and out-of-school suspensions; number of each grade category (A, B, C, D, and F); dummies for ages 14-15, 15-16, and over 17; black and Hispanic dummies; an indicator for having an Individual Educational Program (IEP); a linear grade term; and dummies for having zero, one, two, or three and over arrests of each type. For the one case with missing baseline covariates, we assign a value of zero and include an indicator that the variable is missing.

[18] One slight complication in estimating equation (1) is the possibility that the outcomes of students attending the same school might be correlated. The fact that equation (1) conditions on school fixed effects accounts for within-school correlations across students in mean outcomes. Yet it is still possible that higher order moments are correlated within schools (e.g., school-level variances may be heteroskedastic). Clustering standard errors on schools would account for any remaining correlation across error terms. However, with only 18 clusters, the asymptotic theory on which clustering is based is not applicable. Cameron, Gelbach and Miller (2008) show the wild-t bootstrap performs well with a small number of clusters; as a sensitivity test, we also show p-values from this method.

ITT effect on outcomes divided by the ITT effect on intervention participation rates, varying somewhat due to covariate adjustment. With participation rates of 49 percent for youth assigned to treatment and 5 percent for controls, the IV estimate will be about 2.3 times the ITT.

One complication is that we suspect that there might have been some under-reporting of participation in the after-school programming, which would lead the IV estimate to overstate the effects of participation by under-estimating the number of compliers over whom the ITT impact should be allocated. As a check on how large this over-statement could be, we also present results that assume the participation rate for after-school programming in every school is equal to the rate we see in the school with the highest after-school take-up rate (70 percent). This surely over-states participation rates, and so serves as a lower-bound on the effects of participation.[19]

To help judge the magnitude of our IV estimates, we also estimate the average outcomes of those youth in the control group who would have complied with treatment had they been assigned to treatment – or the "control complier mean" (CCM) (see Katz, Kling & Liebman 2001), which could differ from the overall control mean. Katz, Kling, and Leibman's original formulation of the CCM is in a setting where there is no control crossover (no "always-takers"). If C indicates being a "complier" and Z indicates treatment assignment, the CCM equals:

$$(4) \quad CCM = E(Y|C=1, Z=1) - [E(Y|C=1, Z=1) - E(Y|C=1, Z=0)].$$

The term in brackets is our LATE estimate. However, we must recover the first right-hand-side term, E(Y|C=1, Z=1), since what we observe in the data is the mean outcome for all treatment group participants - a weighted average of the mean outcomes for compliers and always-takers. Let P indicate actual participation and A be an indicator for always-takers. Then:

[19] We randomly select non-participants to reassign as participants until the target "participation rate" is reached. We also re-calculated the IV assuming that within each school, the participation rates were actually the same for after-school and in-school programming, which is usually bounded by the other two approaches (available upon request).

$$(5)\quad E(Y \mid Z=1, P=1) = E(Y \mid Z=1, C=1)\left(1 - \frac{E(A \mid Z=1)}{E(P \mid Z=1)}\right) + E(Y \mid Z=1, A=1)\left(\frac{E(A \mid Z=1)}{E(P \mid Z=1)}\right)$$

To recover E(Y|Z=1, C=1) for the CCM calculation, we can estimate the left-hand side and E(P|Z=1) directly from the data, and use random assignment to replace E(A|Z=1) with E(A|Z=0) and E(Y|Z=1, A=1) with E(Y|Z=0,A=1).[20] In other words, we assume treatment- and control-group always-takers are equivalent on average.

A few final methodological issues that arise in our analysis have to do with whether our results are sensitive to alternative approaches for handling missing outcome data or adjustments to our hypothesis tests for multiple comparisons (they are not), and our approach for examining how impacts may vary across schools and individual program providers as a way to think about the generalizability and scalability of this intervention. We discuss our approach to all three of these issues in Section X below, together with the results of those analyses.

VIII. CRIMINAL BEHAVIOR

Table III shows that the program generates very large reductions in arrests for violent crimes and "other" crimes during the program year (arrests made between 9/09 through 8/10), which are no longer statistically significant in the follow-up year (arrests that occur from 9/10 through 7/11). The top panel shows that during the program year, program participation reduces violent-crime arrests by about 8 per 100 youth, equal to about 44 percent of the CCM (18 per 100 youth). Even the lower-bound estimate for the IV, which assumes a high-end program participation rate, implies a 32 percent reduction. This result is statistically significant, even using the wild-t bootstrap ($p<0.04$). Violent crime, dominated by assaults, is the offense category we would perhaps expect to be most strongly affected by an intervention with a major emphasis

[20] In our case, block randomization means that these equalities should also be conditional on school. In practice, the difference that calculating them conditionally makes is trivial.

on reducing overly automatic angry behavior and mistakes reading other people's intentions. (Violent-crime arrest rates are high relative to the other crime categories in Table III, which may be due partly to the better coverage in the ISP data of juvenile felonies versus misdemeanors).

The top panel's last row shows that program participation reduces the number of arrests for "other" (non-violent, non-property, non-drug) crimes by nearly 12 arrests per 100 youth during the program year, about a 38 percent reduction relative to the CCM (32 per 100). This impact is driven by reductions in weapons offenses, trespassing, and vandalism (each account for about one-quarter of the total effect). Arrests for disorderly conduct or disobeying a police officer, which together account for over a third of all arrests in this "other" category and could in principle have declined simply because youth are now just better able to interact more constructively with law enforcement, appear to be basically unaffected.

The bottom panel shows that for the follow-up year, none of the impact estimates is statistically significant. The impact on "other crimes" is still large in proportional terms, equal to over one-third of the CCM, but is not quite statistically significant (t=1.61). Arrests decline for our sample from year 1 to 2, consistent with declines in overall Chicago crime over this period.[21]

Our results are robust to modifications to our estimation approach (Appendix Table A2). For example, when we lower the probabilistic-match-quality threshold for what counts as a rap-sheet "match," the results are generally similar but slightly attenuated, as we would expect from

[21] For example, the overall rate of Chicago homicides (the best-measured crime) declined by 10% from 2008 to 2009, by 6% from 2009 to 2010, and by another 1% from 2010 to 2011. Declines in youth homicide arrests are even more pronounced. There were 15 homicide arrests to 14-16 year olds in 2008 and just 11 in 2011, and 164 among 17-25 year olds in 2008 and just 91 in 2011 (CPD 2011b). From 2009 to 2010 (the last year for which other crime data are available in Chicago) robbery rates declined by 10% and aggravated assault rates declined by 8.5%, while property-crime arrests were basically unchanged (0.4% drop) (CPD 2011a).

including more false-positive arrests. We also find similar results using a quasi-maximum likelihood Poisson count data model (Wooldridge 1999) rather than OLS. [22]

IX. SCHOOLING RESULTS

Table IV shows that the program improves schooling outcomes during both the program year (AY 2009-10) and the follow-up year (AY 2010-11). We show results for our index followed by each element, all in Z-score form (results in raw units are in Appendix Table A1). The estimated effect of participation on our schooling index during the program year (top panel) is a statistically significant .14SD, with a lower bound of 0.09SD using a conservative adjustment for the potential of attendance under-reporting. The statistical significance of our results is robust to using a wild-t bootstrap to calculate p-values ($p<.034$). The CCM during the program year is positive (0.218), compared to a control mean that is zero by construction, suggesting that youth who are relatively more school-oriented tend to be the ones who choose to participate when offered. The rest of the top panel shows qualitatively similar impacts on each standardized element of our schooling index. Although GPA is the only element significant on its own, the p-values for the other two elements are just barely above the 10 percent level. Using a composite improves our statistical power.

It is worth noting that while the treatment and control mean values for GPA were similar at baseline (Table I), as the first panel of Figure II shows, the variance of the baseline GPA distribution is a bit smaller for the treatment than control group. Specifically, youth assigned to treatment had 0.18 more C's during the pre-program year ($p = 0.06$) and 0.12 fewer A's ($p =$

[22] The QMLE estimates are slightly less precise due to the use of a parsimonious set of covariates to ensure convergence, but the results are almost identical to OLS. Focusing on the main estimate for violent crime, and using only baseline age dummies and the total number of baseline arrests as covariates, the QMLE Poisson coefficient is -0.2011 ($p=.085$). In other words, assignment to the treatment group (the ITT effect) reduces violent crime arrests by about 20 percent. By comparison, the ITT estimate using the same covariates is -0.0328 ($p = 0.052$); a 19.7 percent reduction in arrests relative to the control mean of 0.167.

0.13) than their control counterparts.[23] However, the treatment-control difference in GPA distributions during the program year (shown in the second panel of Figure II) does not appear to be due to this sampling variation. A Kolmogorov-Smirnov test for the equality of distributions, adjusting for school fixed effects, shows that the baseline grade distributions are not significantly different at baseline (p=0.335), but are statistically different the year of the program (p=0.022).

The bottom panel of Table IV shows that the impact on schooling outcomes persists through the follow-up year, and is, if anything, slightly larger than the impact observed during the program year (IV estimates of 0.19SD versus 0.14SD, respectively).

X. ADDITIONAL ROBUSTNESS CHECKS AND EXTENSIONS

This section shows that our results are generally robust to using different approaches to handle missing data on our schooling outcomes and accounting for the number of hypothesis tests we presented in the preceding two sections. We cannot reject the null hypothesis that the effects are the same across treatment arms, and that we cannot reject the null hypothesis that the effects are similar across schools as well – although these are fairly low-powered tests. We also find suggestive evidence that more disadvantaged students may benefit more from the program.

A. Different Approaches to Handling Missing Data on Schooling Outcomes

Table V shows that our results are fairly robust to how we handle missing data on schooling outcomes during the program year (top panel), and the follow-up year (bottom panel), which is perhaps not surprising given that the share of observations with missing data on either the GPA or days-attended variables in our schooling index is nearly identical for the treatment and control groups (10 percent during the program year, and 33 percent during the follow-up year, with no missing data on the school enrollment element of the index by construction).

[23] As discussed below, simulations suggest a substantial probability of observing at least one unbalanced baseline outcome such as this that arises through chance rather than through randomization failure.

Missing data is not an issue we can explore with the "rap sheets" we use to measure arrests, which cannot distinguish between missing data and someone who has just never been arrested.

The first row of Table V reproduces our main results, which follow Kling, Liebman, and Katz (2007) (KLK) and assign the relevant treatment or control group mean to youth with missing values on any element (essentially averaging the results of separate regressions on each of the index elements using just non-missing elements of each index). This approach assumes elements of our outcome index are missing completely at random (MCAR), that is, for reasons uncorrelated with observed or unobserved attributes of youth in the study.[24] MCAR is a testable assumption that seems to fail in our application,[25] perhaps because CPS data on grades and attendance can be missing because youth attend or transfer into particularly low-performing schools with poor record-keeping, transfer to private or suburban schools, or drop out.

The remainder of Table V presents the results of alternative approaches to dealing with missing data that generally yield qualitatively similar results to those from our main approach. The second row shows that the results of using just observations with non-missing values on all elements of the index (list-wise deletion) and controlling for baseline covariates, which also assumes MCAR. Row three again uses complete cases but re-weights the data so the distribution of baseline characteristics in this sample is similar to what we see in the full study sample. This approach assumes that the data are missing at random (MAR), i.e., in ways that are related to youth observable characteristics but not unobserved determinants of outcomes. The results are slightly smaller than our main findings with slightly larger standard errors. The next two rows of

[24] While we do control for baseline covariates, this is not enough to account for correlation between baseline covariates and data missingness, because the ITT or LATE estimates are averages across different cells defined by the baseline covariates. So if some baseline covariate values are over-represented among those observations with missing outcome data, the estimated effect on that sub-sample will be under-represented in the overall estimate.
[25] Regressing an indicator for having non-missing values on all three index elements on baseline covariates produces a global F-statistic of 3.57 ($p<0.0000$) for the program year and 15.52 for the post-program year ($p<0.0000$).

Table V use logical imputation.[26] The last row of each panel presents the results from a multiple imputation (MI) approach with $m = 10$ imputed data sets, which again assumes MAR and yields quite similar estimates (see Appendix D for details).

A final approach we employ to deal with missing data in our CPS schooling outcomes is bounding, motivated by the recognition that even the MAR assumption need not hold. The bounds that we calculate using the trimming procedure from Lee (2009) are still consistent with the idea of large treatment effects on schooling outcomes but, perhaps not surprisingly, have wider confidence intervals than we find for our main estimates, and which now include zero.[27]

B. Multiple Testing Adjustments

As more hypotheses are tested, the probability of at least one false rejection increases. While we have tried to reduce risk of Type I error by minimizing the number of outcomes we examine and by using a large sample size, we use two additional methods to test the robustness of our results to multiple hypothesis testing concerns (see detailed discussion in Anderson 2008).

The first method controls the family-wise error rate (FWER), or the probability that at least one of the true null hypotheses in a family of hypothesis tests is rejected, using a free-step down resampling method to adjust our p-values to account for multiple inference concerns.[28] Our

[26] For our logical imputation we first fill in zeros for all missing grades and attendance information under the extreme assumption that all missing data are due to dropout; in the following row, we set grades and attendance to zero only in those cases where the enrollment variable is zero and the CPS leave codes (which themselves may be subject to some error) suggest the student dropped out (and using the KLK approach otherwise).

[27] In calculating these bounds, we consider an observation non-missing only if all three academic components of the index are non-missing. In year 1, 0.023 more treatment youth than control youth have non-missing indexes; in year 2 the difference rises to 0.046. Trimming off the p^{th} and $1- p^{th}$ quantiles results in a lower bound treatment effect in year 1 of 0.0195 (0.0334) and an upper bound treatment effect of 0.1019 (0.0607) (standard errors in parentheses). For year 2, the analogous bounds are 0.0037 (0.0443) and 0.1613 (0.0607). Using Lee's preferred confidence interval construction, this implies that the true academic treatment effect in year 1 is between -0.035 and 0.1699, and in year 2 is between -0.0692 and 0.2613. However, there are indications that the monotonicity assumption on which this strategy relies may not hold, see Appendix D for discussion.

[28] Specifically, we use a bootstrap resampling technique that simulates data under the null hypothesis (Westfall & Young 1993). Within each permutation, we randomly re-assign treatment and control indicators with replacement and estimate program impacts on all five of our main outcomes (the schooling index and our four main arrest categories). By repeating this procedure 100,000 times, we create an empirical distribution of t-statistics that allows

bootstrap resampling approach suggests we would observe an effect as extreme as the one on the schooling composite by chance only 3.8 percent of the time. So we can confidently reject the null hypothesis that the intervention effects are jointly zero.[29] Given how little we know about improving life outcomes for disadvantaged youth, especially boys, this is a key finding.

While the FWER-adjusted p-value on violent crime arrests (unadjusted p=0.04) is equal to p=0.16, which suggests some caution in interpreting the violent crime results, the FWER control method is somewhat conservative in trading off power in exchange for minimizing the chance of even one type I error. A different approach is to control the false discovery rate (FDR), or the proportion of null-hypothesis rejections that are type I errors (Anderson 2008; Benjamini & Hochberg 1995). For our group of five hypothesis tests, we can set the acceptable expected proportion of type I errors (call this *q,* which is FDR control's p-value analog), then test whether we can reject each null hypothesis at the acceptable *q*- level. Using the two-stage FDR-control procedure from Benjamini, Krieger & Yekutieli (2006),[30] we can reject the null of no violent-crime arrest effects at $q = 0.1$. In other words, as long as we are willing to accept that 10 percent

---

us to compare the actual set of t-statistics we find to what we would have found by chance under the null. We maintain the original sampling frame for each iteration, blocking on schools and assigning the same number of pseudo-treatment and pseudo-control youth as in our original sample. This technique preserves the correlational structure and underlying distributions of our data, providing the adjusted probability we would observe our results by chance given our data and the number of tests we run. Rather than use a single p-value adjustment for all the outcome measures, we use a free step-down procedure to adjust the p-value on each outcome separately. The idea is that once a null hypothesis has been rejected via the bootstrap resampling method, it is removed from the family of hypotheses being tested (thus increasing the power of the remaining tests). We then calculate a new adjusted p-value with the bootstrapped empirical distribution of t-statistics for only the remaining tests, providing a more powerful adjustment than setting all p-values to the same minimum value.

[29] This bootstrap resampling technique also suggests that the one marginally significant baseline imbalance we see in our data is due to chance, not randomization failure. Specifically, the chance of seeing at least one statistically significant treatment-control difference at the 5 percent level out of 19 baseline measures (we see one difference just over that level for number of C's, with $\beta = 0.18$, $p = 0.06$) is 62 percent.

[30] The authors show that this method sharpens the original formulation of FDR control as long as p-values are either independent or positively correlated. We expect that our p-values are positively correlated in this case. The intuition of the original test is as follows: suppose there are M hypotheses, r of which are rejected. Starting with the largest p-value, check if $p < (q*r)/M$. If not, continue to the next largest p-value. Once a p-value satisfies this inequality, reject that null hypothesis and all hypotheses with p-values smaller than that one. This method controls the FDR at level q.

of our rejections will be type I errors in expectation, our violent crime results are robust to adjustments for multiple hypothesis testing.

C. Effects by Treatment Arm

This section shows that the ITT effects across treatment arms appear to be fairly similar to one another. We have focused up to now on pooling the three treatment arms together, in part for improved statistical power, in part because treatment-group cross-over (Figure I) makes it hard to learn about specific mechanisms by comparing effects of different treatment arms, and in part because the coaches involved in after-school programming received BAM training and were encouraged to deploy those skills during the after-school programming. We focus here on comparing the ITT across arms rather than the IV because instrumenting for "participation" by arm is complicated by the differential under-reporting of attendance across arms.[31]

Table VI shows that the ITT effects are fairly similar across treatment arms; we cannot reject the null that they are the same. A different approach (Appendix Table A3) assumes that the effects of in-school and after-school programming are additive and regress outcomes against an indicator for being assigned to in-school and after-school, so that youth assigned to both have both indicators turned on. Again we cannot reject the null that the coefficients are the same.

The similarity of these ITT effects across treatment arms is another reason that we suspect that the recorded participation rate for the after-school-only group (21 percent) is too low. Since around half the youth in the other groups participated, for the 21 percent participation rate to be correct, the after-school programming would have to be far more effective per participant than the effects of either in-school programming alone or even than the combination of in-school and after-school programming.

[31] When we instrument for participation in each activity type with the three treatment-arm dummies, we cannot reject that the effects of the in-school and after-school activities are the same. This is true regardless of whether we use the participation data as-is or our bounding approach that adjusts for after-school under-reporting.

D. Treatment heterogeneity

When we interact different baseline characteristics of youth with treatment assignment, we generally find few statistically significant differences across identifiable sub-groups of youth in estimated impacts. One possible exception is that schooling impacts might be relatively larger for youth who were towards the bottom of the baseline GPA distribution (below a D), who experienced an impact on the schooling index of 0.11 standard deviations (p=0.027) compared to a statistically insignificant 0.026SD change (p=0.314) for students with higher baseline GPAs (Appendix Table A4). The table also shows that the program year's decrease in violent crime appears to be concentrated among those who had not yet been arrested for a violent crime at baseline (β= -0.0818, p = 0.050, versus for those who had already been arrested for a violent crime at baseline, β = 0.030, p = 0.417). We should, however, be cautious interpreting these findings, given the number of hypothesis tests involved in testing all the interaction effects.

A different concern is that the effectiveness of programs that seek to develop human capital may vary considerably by the individual program provider, or across settings, so that scaling-up the results might be challenging. To explore this possibility – at least to the extent possible within a single study – we take advantage of the fact that our intervention was carried out by 13 different individuals for the in-school programming and 21 different after-school coaches, across 18 CPS schools that differ by racial and ethnic composition, poverty rates, academic achievement, and a variety of other characteristics. We find little statistical evidence that treatment effects differ across schools using either ordinary least squares or hierarchical linear modeling (which some researchers prefer because of the efficiency gain that comes from precision-weighting the school blocks); see Appendix E for details. Given that we have only 18 schools, it should be noted that this is a fairly low-power test. With that qualification in mind,

and recognizing that replication would provide the strongest evidence about external validity, we cannot reject the null hypothesis that effects are similar on average across schools.[32]

XI. MECHANISMS

Given our research design and reliance on administrative data, our ability to isolate the key mechanisms of action behind the sizable behavioral impacts we document here is somewhat limited. We can rule out after-school incapacitation as the sole reason behind the observed arrest impacts. The available survey data we have from CPS, while imperfect, provide at least suggestive evidence that the intervention changes self-reported persistence and peer relationships or conflict resolution. Our data have little to say directly about other candidate channels such as a generic mentoring effect (or exercise or program-induced change in peers), aside from noting that other interventions with these elements do not appear to be as successful as this one.

The arrest data are not consistent with the idea that crime impacts are due solely to the program incapacitating youth during high-risk after school hours. We observe the date of each arrest and also know the dates of each after-school session. We find that the estimated effect is not concentrated on days in which after-school programming is concentrated, which is not consistent with incapacitation being the major mechanism behind arrest impacts.[33]

A second possible mechanism that may be at work is the role of mentorship. It may be that having a positive, pro-social adult consistently checking in with, advising, and advocating for a youth can improve his schooling and crime outcomes. This is possible, especially during the program year, although the fact that the observed schooling gains persist through the year

[32] We find that the treatment effect's variance is quite small: 0.00005 for the program year academic treatment effect, 0.0002 for the follow-up year academic treatment effect, and 0.0003 for the violent crime arrest effect.
[33] The ITT effect on an indicator for any violent-crime arrest during days when after-school programming is not offered is β = -0.0217 (se=0.0103), p = 0.035, CM 0.046, vs. days after-school programming is offered, β = -0.0061 (0.0076), p = 0.420, CM = 0.094). These estimates do not adjust for the larger number of non-programming days.

after the program, when students were no longer interacting with those adults, suggests increased supervision and mentoring might not be the only mechanism at work.

The best available outside evidence on this mechanism probably comes from the randomized experimental studies of the Big Brothers / Big Sisters (BB/BS) mentoring program, although these rely on self-reported survey data and so may confound behavioral effects with effects on willingness of youth to admit bad behavior. In any case, the findings are somewhat mixed. An initial RCT of BB/BS community-based mentoring for children ages 10-16 found evidence of beneficial effects on schooling outcomes like GPA and attendance, and some behavioral measures such as drug and alcohol use or hitting someone else (but not theft or property damage) (Grossman & Tierney 1998). The BB/BS study finds few effects when examining our study sample (minority males). A more recent study of BB/BS mentoring done within schools with children 10-16 (Herrera, et al. 2011) found effects on some schooling outcomes during the program year that did not persist to the follow-up year, and found no statistically significant effects on out-of-school behavioral outcomes.

A more subtle way in which the intervention could have increased supervision and mentoring of youth is by reducing school-switching. Table VII shows that during the program year, the likelihood of changing schools within the CPS system was about 50 percent lower for program participants. Reduced school mobility could have improved youth outcomes both because changing schools is directly disruptive, and because youth who stay in the same school may be more likely to have or develop relationships with school staff. Table VII also shows participants are about half as likely to have attended a CPS school in the criminal justice system (juvenile detention or prison, or adult jail or prison) the year after the program. If incarceration has adverse effects on youth, this impact could be a mediator for lasting changes in school

engagement. This could also be a behavioral outcome in its own right, or simply a mechanical (but substantively important) after-effect of the reduction in violent-crime arrests during the program year itself, since the decision about whether to detain a youth after arrest in Cook County is based on a formula that includes a youth's prior criminal record as one input.

Our final source of information about potential mechanisms, including possible effects on judgment and decision-making, comes from surveys completed by CPS students in spring 2011, the end of the year *after* our intervention. These on-going, bi-annual surveys are carried out over the web by the Consortium on Chicago School Research and are designed to measure student perceptions of themselves and their school environment.[34] The two most relevant measures for present purposes are "emotional health," which captures some combination of conflict resolution and social information processing, and persistence, which could be thought of as measuring either some social-cognitive skill ("grit") or the tendency for youth to make a particular judgment and decision-making error (catastrophizing) that leads them to desist after experiencing some failure (Figure III). One limitation of these surveys is that the response rate for our sample is low, and is slightly different for treatment versus control groups (42 versus 38 percent, $p<.05$).

With these limitations in mind, Table VIII shows that the effect of participation on a Z-scored index of our measures of perseverance and emotional health is equal to 0.13 standard deviations.[35] While this estimate is somewhat imprecise ($p=0.180$), it is at least in a direction that is consistent with changes in some underlying decision-making or non-academic skills may be one mechanism at work here. Suggestive evidence that this estimate is not merely an artifact of

[34] The 30-minute survey is designed to address a number of questions regarding school culture and climate. CCSR has administered surveys to CPS teachers, students, and principals for two decades. In Spring 2011, surveys were received from ~146,000 students in more than 600 schools. All students in grades 6-12 and all teachers were asked to participate. The student web survey was administered within each school during school hours (CCSR Support Center 2012), with each response registered on a Likert scale. CCSR used Rasch analysis on individual survey items. We standardized these measures into standard deviation units based on the observed distribution within the control group.

[35] This index was formed in the same way as our main academic index.

low or differential survey response rates comes from the fact that we do not see similar positive impacts on measures that should not be affected our intervention, such as *academic press* (how challenging students find courses and teachers) and *course clarity*.

XII. BENEFIT-COST ANALYSIS

This section presents our attempts to measure the monetized value of the benefits and costs of the program. The cost side is driven by tangible costs of delivering the program to participating youth, and so is easy to measure from budget information from the program providers – around $1,100 per participant.[36] Monetizing the benefits of the program, which includes intangible benefits such as improved quality of life from less street crime to the public, is more complicated. We summarize the results of our attempts here, with details in Appendix F.

The results presented in Table IX suggest that the value to society from reductions in criminal behavior during our study period, concentrated during the program year itself, may generate benefit-cost ratios ranging from 5:1 up to 30:1. The smallest estimate ($5,309) comes from using our lower-bound IV and lower-bound monetary valuations for homicide, which wind up driving cost-of-crime estimates (Kling, Ludwig & Katz 2005). The upper bound ($33,262) comes from assuming participation data are accurate and using willingness-to-pay estimates for the costs of crime. Our standard errors understate the true uncertainty here because they ignore the conceptual uncertainty associated with monetizing the value of reductions in crime.

Because homicide is so much more costly than all other offenses, readers might wonder to what degree these benefit-cost figures are driven by impacts on this costly but relatively rare outcome. If we exclude homicide from the analysis altogether, the estimate for the IV effect on the total number of violent crimes per youth is perhaps unsurprisingly similar to our main result

[36] Since program costs are most often calculated on a per-participant rather than per-random assignee basis, we use the LATE impact estimates to calculate benefits per participant. We could equivalently use the ITT impact estimates, but rescale the costs to be per random-assignee.

presented above. Our lower- and upper-bound estimates for the social benefits from reductions in criminal behavior by youth (the top row of Table IX) now equal $2,300 and $12,150, respectively, and are much more precisely estimated than are our estimates with homicide included (with p-values equal to .04 and .01, respectively). Put differently, the benefit-cost ratio is at least 2:1 just from the social-cost impacts from offenses aside from homicide.

The benefits might be much higher still if the impacts on schooling outcomes that we see during our study period – both in the program year and follow-up year – lead to increased high school graduation. To approximate what these benefits could be, we use a study of district-wide longitudinal data that reports how schooling outcomes during grade 9 correlate with later graduation rates (Allensworth & Easton 2007). We use these estimates to approximate the change in graduation rate associated with the improvement in GPA seen in our treatment group, then multiply those graduation rates by estimates for the benefits of increased high school graduation. This extrapolation obviously requires strong assumptions. The magnitude is nonetheless notable: we forecast that the changes in GPA caused by the program could translate into increases in graduation rates between 3 and 10 percentage points, or 7 to 22 percent relative to control complier baseline rates, which could translate into benefits to society that range from about $38,000 per participant up to as much as $84,000.

This analysis also highlights the difference across disciplines in how to think about the value of policy interventions. For example, the influential Blueprints for Violence Prevention program requires an intervention to have impacts on crime that persist in order to be a "model program." In contrast the standard argument by economists would be that for purposes of guiding public policy, what matters is not the persistence of an impact but rather the whether impacts of whatever duration generate benefits larger than costs (as is the case here).

XIII. CONCLUSION

This paper presents results from a large-scale randomized experimental study in Chicago with economically disadvantaged male youth, which suggest that behavior for this population may be much more elastic than previous research suggests. Participation reduces violent-crime arrest by 8 per 100 participants (44 percent) and arrests for "other" offenses by 11.5 per 100 participants (36 percent), driven by declines in vandalism, trespassing, and weapons offenses. These impacts occur on the types of offenses we might expect to be most strongly affected by an intervention that places a heavy emphasis on reducing automatic, angry behavior, and common biased beliefs like hostile attribution bias (over-attribution of malevolent intent to others). While the crime impacts do not persist, impacts on schooling outcomes do, with gains that we estimate could translate into higher graduation rates of 3 to 10 percentage points (7-22 percent).

While our reliance on administrative data limit our ability to empirically isolate mechanisms of action, these impacts seem to be much larger than what we see from other interventions that include shared ingredients like mentoring or after-school programming. We think there is at least a suggestive case to be made here that the novel ingredient in this intervention – CBT designed to reduce judgment and decision-making errors – may be an important mechanism of action.

Why does this intervention have lasting effects on schooling but not crime? Logically, there are two possibilities. One is that there are at least two latent factors affected by the intervention, one of which matters more for schooling and the other for crime, and impacts on the one factor persist longer than on the second. The other possibility is that there is a single underlying factor responsible for changes in both schooling and crime, with a more extreme threshold for violence than for school disengagement or dropout (given the former is less

common than the latter). We would see less persistence in crime impacts if the intervention's effects on this underling factor persists less for the highest-risk youth (as seems to be the case with young children in Head Start, at least for achievement test scores; see Currie & Thomas 1995; Deming 2009), so the shape of the treatment-group distribution changes over time. Knowing more about the causes and remedies for program "fade-out" remains an open question for our study and for social policy more generally.

As with all randomized experiments, there is always some question about the degree to which these impacts generalize to other samples and settings. Because our study was carried out with large numbers of disadvantaged male youth from distressed areas of Chicago, it is closer to an "effectiveness trial" (testing a program at scale) than an "efficacy trial" of a model (or "hothouse") program. The intervention we study should, in principle, lend itself to further scale-up, given that it is manualized, and given that estimated benefit-cost ratios for the large-scale implementation we study range from 5:1 to 30:1. However, because keeping youth engaged is so central (and so difficult), selecting the right people as providers may be particularly important.

Given the sizable impacts and benefit-cost ratios we estimate here, for a population (disadvantaged youth) for which there is currently not a surplus of successful intervention examples, replicating these results would seem to be a priority for future research. In that spirit we are encouraged that a separate CBT experiment we carried out in the Cook County juvenile temporary detention center also seems to have positive early findings on the risk of readmission (Heller, Guryan & Ludwig 2012).

What is perhaps most surprising about these findings is the size of the gains in schooling outcomes (which could translate into increased graduation rates of 7-22%) and observed reductions in violent-crime arrests (44%) given the relatively limited number of one-or-two-hour

sessions participants attended (about 13) and the low cost of the intervention ($1,100 per participant). The behavioral outcomes of a study sample (disadvantaged male youth) that have been so hard to help through other interventions appear to be remarkably elastic to even fairly modest investments in a program that includes CBT-based efforts to remediate common, predictable judgment and decision-making errors. Yet most of the $550 billion the U.S. spends on our most important socializing institution – our K-12 public schools (U.S. Census Bureau 2010) – is devoted to developing academic skills, at least after the first few years. Given how little attention is currently devoted to addressing non-academic factors that affect long-term outcomes of at-risk youth, there may be substantial returns to society from expanding investments in this area.

Appendix A:
Review of previous intervention studies

We begin this appendix by reviewing what is known from previous studies about the impact of cognitive behavioral therapy (CBT) and related interventions on outcomes of youth. We then expand the lens to studies that examine both older and younger populations. We conclude by noting that remarkably little is known about how to improve schooling outcomes and reduce criminal involvement among at-risk teens even when we look across the entire intervention literature, not limited to any specific intervention strategy.

1. Previous assessments of cognitive-behavioral interventions for youth

Previous meta-analyses of targeted, low-cost approaches that fall under the general rubric of Cognitive Behavioral Therapy (CBT) claim that the available empirical evidence supports the value of this intervention. However, our careful inspection of individual studies in this literature suggests that claim rests mostly on findings from non-experimental studies that may confound the effects of CBT interventions with the effects of selection of systematically different types of youth into programming or comparison conditions. Our examination of the modest number of previous randomized experiments that have been carried out suggests more mixed findings in support of CBT, if these results are taken at face value. But even many of the experiments have important methodological limitations as well, which leads us to conclude that the current studies are not very informative about the effects of CBT on those youth outcomes of greatest policy concern, such as high school dropout or violence involvement.

Up through the 1970s, most psychological interventions focused on helping people to identify and process conflicts and traumas in their pasts. Traditional psychodynamic approaches view presenting symptoms as reflecting more fundamental underlying difficulties which must be addressed before the presenting symptoms could be genuinely relieved (Walker & Bright 2009, p. 179). CBT is more pragmatic in its objectives. A key innovation of CBT was to recognize that the effects of past experience on current problematic symptoms are mediated through problematic, often automatic thoughts, and that focusing more directly on those mediating thoughts can lead to greater short-term relief of symptoms. Compared to traditional psychological approaches, CBT is also more directive, pursuing specific goals such as symptom relief or behavior change, and more structured, focused on concrete problems and their solutions.

CBT is a broad label, which encompasses a family of problem-focused treatments (e.g., Rational Emotive Behavior Therapy, Rational Behavior Therapy, Rational Living Therapy, Cognitive Therapy, Dialectic Behavior Therapy) that follow similar guiding principles and seek to address related emotional and behavioral problems. CBT includes a variety of techniques to help individuals "identify, monitor, challenge, and change their thoughts and behaviour" (Walker & Bright 2009, p. 179). CBT's motivating principles include the belief that maladaptive thoughts are key antecedents to problematic emotions and behaviors. When CBT is successful, individuals learn more effective patterns of thinking and relating to their environments. Individuals also learn new strategies to regulate automatic or impulsive behaviors. By helping people to think more realistically and effectively, interventions can provide symptomatic relief while ameliorating problematic behaviors.

Specific CBT intervention strategies vary, though common elements distinguish CBT from other behavioral interventions (Walker & Bright 2009, p. 179). CBT requires patients' or clients' active participation in the treatment process. Treatment providers frequently employ individual or group exercises, role-playing, or individual storytelling to make CBT an active collaboration between treatment providers and those seeking to benefit from treatment intervention. In practice, CBT participants are often ambivalent regarding deeply-rooted problematic behaviors, and are often ambivalent regarding continued participation and engagement in the treatment itself. Motivational components are therefore especially important for any successful intervention. CBT is also time-limited. Relatively brief interventions are expected to produce tangible benefits. Most CBT interventions are relatively short-duration (generally 16-24 contact hours).

CBT has been shown to be effective in providing symptomatic relief for specific psychiatric disorders such as depression (Birmaher, et al. 2000; Brent, Holder & Kolko 1997; Clarke, Hops & Lewinsohn 1992; Rohde, et al. 2004; Wood, Harrington & Moore 1996), anxiety disorders (Barrett, et al. 2001; In-Albon & Schneider 2007; Kendall & Wilcox 1980; Kendall, et al. 1990), intermittent explosive disorder (McCloskey, et al. 2008), conduct problems (Kazdin 1995; Kendall & Wilcox 1980; Kendall, et al. 1990; Koegl, et al. 2008), attention deficit hyperactivity disorder (Toplak, et al. 2008), and emotional dysregulation among severely disordered people (Koerner & Linehan 2000; Linehan, et al. 1999). CBT has also been found to help treat problems like chronic pain (McCracken & Turk 2002), medication adherence (Parsons, et al. 2007), adolescent substance use problems (Waldron & Turner 2008; Waldron & Kaminer 2004), and stress management (Antoni, et al. 2000; Gaab, et al. 2003).

Growing practitioner interest has led to attempts to use CBT to change other youth problem behaviors as well. CBT interventions have tried to reduce youth problem behavior by helping youth to reduce automatic impulses and aggressive behavior, through for example teaching relaxation techniques. CBT is also used in efforts to help youth broaden their perspectives in making choices (such as helping youth better consider the consequences of their actions for others), and measures to address specific cognitive distortions such as hostile attribution bias (assuming others have malicious intent). Interventions may also develop and practice specific problem-solving techniques, including techniques for conflict resolution.

Several meta-analytic reviews conclude that CBT *might* be a very effective (and very cost-effective) way to reduce crime and delinquency among both adults and juveniles (Drake, Aos & Miller 2009; Greenwood 2008; Landenberger & Lipsey 2005; Lipsey 2009; Lipsey & Cullen 2007). For example, Drake, Aos, and Miller (2009) conclude that "the net value of the average evidence-based cognitive behavioral program for adult offenders is $15,361 per offender." A Campbell Systematic Review by Lipsey, Landenberger, and Wilson (2007) noted many limitations of existing research, but also reach a favorable overall assessment: "Research to date leaves little doubt that CBT is capable of producing significant reductions in the recidivism of even high risk offenders under favorable conditions."

Yet the empirical support for the most optimistic of these claims comes largely from non-experimental studies that are susceptible to selection bias. Those youth or adults who select into

CBT programs may be systematically different from those who did not volunteer, in which case non-experimental studies may confound the causal effects of the programs with those of hard-to-measure individual attributes associated with program selection. While the meta-analyses use statistical tests to gauge whether the presence of non-experimental studies might have skewed their results, and tend to find few statistically significant correlations between specific study-design features and effect sizes, statistical power for this sort of exercise is typically modest.

Unfortunately, randomized experiments are so rare in this area that the highly-regarded Campbell Collaboration – which is dedicated to synthesizing rigorous empirical research and promoting evidence-based policy – concluded "it is unrealistic to restrict systematic reviews in [this] field to randomized experimental studies, however superior they may be, because so few exist" (Greenwood 2008, p. 289). A discomfiting proportion of the experiments included in meta-analyses are technical reports, chapters, or doctoral dissertations rather than published articles in the peer-review literature. Those few experiments that are carried out also tend to focus on small-scale model programs, rather than at-scale programs – that is, they mostly correspond to what medical researchers call "efficacy trials" rather than large-scale "effectiveness trials."[37] And even many of the small-scale tests have important study limitations.

2. Re-examining experimental studies of cognitive behavior interventions for youth

We performed our own careful examination of every randomized trial of a CBT intervention, or of arguably related programs to promote social-cognitive skills or socio-emotional learning. Our sample frame for identifying individual studies was to include every experiment that was included in several particularly influential meta-analyses. To reduce the risk of missing relevant high-quality randomized experiments, we also included any experiment testing a CBT, social-cognitive, or social-emotional learning intervention that was considered to be a high-quality RCT by one of several widely-used research aggregators. Specifically, we included all studies that were either:

1. Included in the review of the entire literature on crime-prevention carried out by the Washington State Institute for Public Policy (Aos, Miller & Drake 2006; Lee, et al. 2012)
2. Included in the review of the literature on cognitive-behavioral programs for criminal offenders by Landenberger and Lipsey (2005)
3. Included in the review of the social and emotional learning literature by Durlak, Weissberg, Dymnicki, Taylor and Schellinger (2011).
4. Was rated a "Top Tier" or "Near Top Tier" intervention by the Coalition for Evidence-Based Policy (www.coalition4evidence.org).
5. Was rated a "Level 1" intervention by FindYouthInfo.gov, which was created by the federal government's Interagency Working Group on Youth Programs

[37] In similar fashion, Lipsey, Landenberger, and Wilson (2007) report: "Of the 58 studies that met the inclusion criteria for this review, only 19 used random assignment designs and, of those, only 13 maintained sufficiently low attrition from outcome measurement to yield results with high internal validity. Moreover, only six of the random assignment studies were conducted on "real world" CBT practice; the others were research and demonstration projects. The amount of high quality research on CBT in representative correctional practice is not yet large enough to determine whether the impressive effects on recidivism found in this meta-analysis can be routinely attained under everyday circumstances" (p.58).

6. Was rated "Effective" by CrimeSolutions.gov, sponsored by the U.S. Department of Justice's Office of Justice Programs
7. Was rated as a "Model Program" by the Blueprints for Violence Prevention (www.colorado.edu/cspv/blueprints), a widely-cited resource established by the University of Colorado to "identify truly outstanding violence and drug prevention programs that meet a high scientific standard of effectiveness."
8. Met the evidence standards of the U.S. Department of Education's What Works Clearinghouse (ies.ed.gov/ncee/wwc/). In addition, we included four other valuable studies that met the What Works Clearinghouse standards "with reservations."

These searches identified 27 studies that focus on youth age 13-18, which, taken at face value, suggest mixed results. Twelve of the 27 studies find beneficial, statistically significant effects; 15 of the 27 find no statistically significant results. Close inspection of the 27 studies suggest that this pattern of mixed results may be less informative than it first appears. Many of the cited studies display important limitations that limit internal or external validity.

Fifteen of the 27 studies displayed faulty randomization, as suggested by either the description of how the authors tried to carry out random assignment, or by evidence of imbalance in baseline characteristics for the "randomized" treatment and control groups. An additional seven exhibited attrition rates of at least 20 percent, or marked differences in the study attrition rates between treatment and control groups. Nine of the 27 studies rely upon self-reported outcomes. We know from prior studies that student self-reports regarding crime and other stigmatized behaviors are susceptible to widespread under-reporting (Kling, Ludwig & Katz 2005). Systematic differences between treatment and control groups in under-reporting of anti-social behaviors is a particular problem for any intervention that develops relationships between program providers and participants, since the latter may then wish to avoid disappointing the former by not confessing to undesirable behaviors (that is, program participation may affect the degree of self-presentation bias on self-reported survey responses).

Sample size (and thus limited statistical power) is also a substantial issue. Seven of the twelve successfully randomized studies involved fewer than 100 individuals in the treatment group. (Three of these seven involved treatment groups of less than fifty.) Small sample sizes are often adequate when exploring common outcomes. Larger samples are required when one explores relative rare outcomes such as violent offending. Even when sample sizes are slightly larger, some important studies yield suggestive findings that fail to reach statistical significance due to low power.

Fourteen of the 27 studies concern youth in specialized juvenile justice or mental health settings. These studies are thus less pertinent to understanding the degree to which this sort of intervention approach can improve youth outcomes at large scale when delivered in community settings, such as through the public schools as in the program we study here. Fully 21 of the 27 studies experienced limitations to randomization, attrition difficulties, or relied on self-reported outcomes. Out of the remaining six studies, only two (Armstrong 2003; Dynarski, et al. 1998) included treatment groups exceeding 100 individuals.

Dynarski et al. (1998) avoided many threats to internal validity common in this research literature. These authors analyzed an RCT of the "Twelve Together" peer support and mentoring program for middle- and high-school students in Chula Vista, California (WWC Intervention Report 2007). Like the current intervention, Twelve Together included weekly peer discussion groups involving roughly twelve participants and an adult facilitator, the latter often a college student. The program also included homework assistance, college trips, and an annual weekend retreat. The trial met WWC evidence standards "with reservations," because treatment-control difference in survey response rates (92% vs. 86%, respectively) exceeded the five percentage-point differential attrition threshold used in WWC reviews of school dropout.

At the end of three-year follow-up, Dynarski and colleagues found that 8% of the treatment group had dropped out of school, versus 13% of controls. Yet the implied effect size of 0.33SD failed to reach statistical significance given the constraints on statistical power in the study – the sample subject to random assignment was just 219. These authors also found no statistically significant benefits in other domains, such as highest grades completed, days absent, dropout, or school disciplinary problems (Dynarski, et al. 1998).

Armstrong (2003) was identified as a high-quality study by the careful literature review carried out by Aos, Miller, and Drake (2006). This study provides a clinical trial of Moral Reconation Therapy (RCT). This experiment randomly assigned a total of 256 juveniles within a Maryland detention center to treatment (N=135) or a control group (N=121). The main results in the paper come from analyzing a sample that excludes the N=19 youth assigned to the treatment group who did not actually receive the treatment because they refused, or could not speak English, or were released from the facility, as well as the N=25 control group youth who wound up receiving the program, and so does not fully represent what one might think of as "best-practice" for analyzing data from a randomized experiment.

Armstrong reports that the experiment also carried out a more standard intention-to-treat (ITT) analysis and the "results of the two sets of analyses were not different" (p. 676), but the ITT point estimates and confidence intervals are not presented in the paper itself. Overall recidivism rates are not different between treatment and control groups; in terms of the number of days to re-arrest, the treatment group has a higher mean (307 vs. 295) and median (258 vs. 228). Given the total number of juveniles assigned to the treatment and control groups in this study, we have some concern about whether the study has statistical power to detect effects large enough to be meaningful from the perspective of a benefit-cost analysis. Moreover, there may have been a problem with randomization: the proportion of youth who are African-American was much higher in the treatment vs. control group (67% versus 48%).[38]

[38] Two additional studies are worth mention for the programs' similarities to BAM. The ALAS (Spanish for "wings") intervention, despite its small sample size, bears clear similarities in structure and curricular content to those of the BAM Sports Edition intervention. Although the study involved a treatment group of only 46 students, treatment and control groups were successfully randomized, with low attrition and administrative data to avoid common pitfalls of student self-reports.

ALAS served students identified to be at-risk due to academic or behavioral difficulties. Each participant was assigned a counselor, who monitored the student's progress, communicated with parents and teachers, and ensured that ALAS services were delivered. The ALAS program includes intensive attendance monitoring, ten weeks of instruction on problem-solving skills using the ALAS Resilience Builder curriculum (WWS 2006).

Only two of the secondary prevention trials identified statistically significant outcome differences between treatment and control groups. Each of these studies displayed at least one significant study limitation.[39] Only two trials included youth over the age of fifteen, thus excluding the peak years of youth criminal offending.

Teachers provided regular feedback to students through program mentors. Families also received training in parent-child problem-solving and related subjects.

Larson and Rumberger (1995) analyzed a sample of 94 students in the Los Angeles Unified School District. These students had participated in ALAS since the beginning of 7th grade, and were first evaluated at the end of 9th grade, and again evaluated at the end of 11th grade. These authors found statistically significant improvements in two important measures: Student school enrollment at the end of 9th grade (98% within the ALAS group vs. 83% among controls), and being "on track" to graduate at that same point (72% vs. 53%). Differences in enrollment and on-track status at the end of 11th grade continued to favor the treatment group (75% vs. 67%, and 33% vs. 26%, respectively). However these notable differences were no longer statistically significant given the small sample.

[39] The Farrell, Meyer, and White (2001) analysis of the Responding in Peaceful and Positive Ways (RIPP) intervention deserves mention given its similarities to the BAM intervention studied in this paper. Randomized at the classroom level, this study employed administrative records to examine one-year follow-up of students' violent behavior in middle school. Within a mixed pattern of findings, these authors found significantly lower rates of in-school suspensions among male RIPP participants. As with other studies, the RIPP evaluation appears to have experienced high non-random attrition rates. Moreover, attrited students were older, had lower grade point averages, lower attendance, and more out of school suspensions than did students who remained in the study.

Farrell, Meyer, Sullivan, and Kung (2003) performed a similar trial, relying on students' self-reported data of recent violent behavior. These authors found statistically significant differences in outcome between treatment and control groups. However, these results may have been influenced by high attrition rates. On average, attrited students were older, had lower grade point averages (though the difference was not significant), and were less likely to come from two-parent households than other participants.

Orpinas et al. (2000) performed a large intervention trial, randomized at the school level, which sought to reduce middle-school students' aggressive behavior as defined by the Youth Risk Behavior Survey. The study relied upon self-reported outcomes, with no demonstration of treatment-control balance at baseline. Participants displayed a 21.5% attrition rate, with more aggressive students more likely to exist the study sample.

Patton et al. (2006) performed a school-randomized trial in which self-reported anti-social behaviors among 8th graders were compared within 12 treatment and 14 control schools. Sample schools were not shown to be balanced at baseline. Six schools dropped out after being selected and were not included in analysis; one school stopped participating during intervention and was excluded from final analysis. The cross-sectional study design precluded analysis of how particular individuals responded to treatment. Finally, between 19 and 34% of students were not surveyed in the evaluation.

Skye (2001) performed an innovative classroom-randomized trial for high school students. This intervention sought to reduce risk of violence as measured by student self-reports on the Eruptive Violence scale. Treatment and control groups were not balanced at baseline. Moreover, reliance on student self-reports, limited statistical power, and unreported attrition rates provide important limitations.

Harrington, et al. (2001) analysis of the "All Stars" character development program provided another informative school-randomized trial of interventions for middle-school students. These authors examined students' self-reported violence towards other persons within matched pairs of treatment and control schools based on demographics and the receipt of free/reduced lunch. This study's reliance on self-report outcome data and its high sample attrition (27.7%) again provide important limitations.

A Norwegian study by Gundersen and collaborators (2006) provide another pertinent analysis of Aggression Replacement Training. Small sample size—the control group was only eighteen subjects--and compromised randomization hinder interpretation of study results.

Finally, Simons-Morton and colleagues (2005) examined the effectiveness of the multi-faceted "Going Places" intervention, which was designed to address a broad array of adolescent problem behaviors. Seven middle schools were randomized to intervention or comparison conditions and students in two successive cohorts (n = 1484) of students. The study relied on student self-reports. It was also hindered by 37 percent attrition rate.

3. CBT interventions with adults

Expanding our lens to also include CBT experiments with adults, not just teens, does not greatly change the picture about what we can learn from existing research about the effects of CBT or related interventions on behavioral outcomes of key policy concern.

Consider, for example, the two CBT experiments with adults that the literature review by Aos and colleagues (2006) think are of the highest quality: both still have important limitations. Van Voorhis et al. (2004) assess the effects of Reasoning and Rehabilitation (R&R) provided to 468 randomly assigned adult parolees in Georgia, with an average age of 30 and generally fairly extensive prior records. (Because their paper provides an excellent critical review of previous studies of R&R, we do not replicate that literature review here). The treatment group typically has lower rates of adverse follow-up outcomes, approximately 8-10% of the control mean, for outcomes such as 9 month re-arrest rates (38% vs. 42%) and 30-month prison re-admission rates (43% vs. 47%).

While these treatment-control differences are not statistically significant, this could be due to limited statistical power. A total of 243 parolees were assigned to the treatment. Sixty percent of those assigned to treatment completed the program. Outcomes through 30 months were available for only around two-thirds of the sample. The resulting confidence intervals on program effectiveness were rather wide. Within the preferred logistic regression, the 95% confidence interval on the odds ratio for rearrest/parole revocation ranged from 0.62 to 1.17. In similar fashion, the 95% confidence interval on the odds ratio for returning to prison ranged from 0.67 to 1.17. Such limited statistical power is quite worrisome in a policy sense. A decline in criminal offending of 8-10% – if real – would (given the costs of crime; see Ludwig 2006) be ample for the intervention to pass a benefit-cost test (Drake, Aos & Miller 2009).

Aos and colleagues identify one other randomized experiment for adults, Ortmann (2000), that they considered to be of highest quality. This study reports treatment-control differences in recidivism equal to roughly 8-10% of the control mean, but which are not statistically significant. The Ortmann study has even less statistical power than that of Van Voorhis *et al*., having enrolled a total of just 111 prisoners (in Germany) in the treatment group.

4. Related interventions among children

Previous research shows that early childhood interventions such as Perry Preschool, Abecedarian, Head Start, and Nurse/Family Partnership, which provide a mix of academic support, parenting training, and other social services between the pre-natal period and age 5, have long-term effects on educational attainment, employment and earnings, and in some cases, crime--*despite fade-out of impacts on IQ or achievement test scores* (Campbell, et al. 2002; Currie & Thomas 1995; Deming 2009; Garces, Thomas & Currie 2002; Lochner 2011; Ludwig & Miller 2007; Olds, et al. 1999; Schweinhart, et al. 2005). Through process of elimination, researchers have inferred that effects on non-academic factors must be the key mediating mechanism for the long-term effects of these interventions. Indirect proxies for "non-cognitive skills" such as teacher-reported behavior and mood are interpreted as support for this hypothesis

(Heckman, et al. 2010). Yet few studies include good direct measures of these skills; these non-academic factors essentially play the role of social policy dark matter.

A few randomized experiments have tried to change non-academic factors among elementary school children as well. The Fast Track intervention worked with children starting in grade 1 and lasting through high school (Conduct Problems Prevention Research Group 2011). Children in elementary school (grades 1-5) were provided with weekly sessions to enhance social-cognitive skills; the program also provided tutoring to children, parent-training groups, and home visits to work on parenting practices. Intensity of the intervention was reduced somewhat during middle and high school. Follow-up studies found the program during elementary school did indeed strengthen social-cognitive skills, developmentally appropriate parenting practices, and child behavior. However, by high school, the intervention no longer had a statistically significant impact on the full study sample; there were signs of impacts on the highest-risk sub-group, but only on outcomes measured by parent report, not child self-report.

An intervention by Hudley and Graham (1993) focused on addressing "hostile intention attribution bias" (a social information processing problem in which people have the tendency to assume malevolent intent by others) by randomizing a sample of 72 African-American elementary school boys (ages 10-12) screened for problems with aggression. Boys were randomized to an intervention that addressed hostile attribution bias, a different program not focused on addressing hostile attribution bias, included to identify any generic program-participation effect, and a control group. At four-month follow up, the study found some impact of the intervention on how boys interpreted the intention of others in study scenarios, some impact on teacher reports of aggression, and no detectable impacts on disciplinary referrals to the office at school.

5. Interventions to reduce dropout and delinquency / youth violence more generally

It is always possible that our review of the literature might have missed some key studies, or that we are being too negative about the quality of evidence in this area. Some support for our critical interpretation arises from the fact that so few intervention strategies to remediate social-cognitive skill deficits meet criteria for top-tier evidence-based programs by organizations specifically devoted to critically assessing existing research evidence. For example, the U.S. Department of Education's What Works Clearinghouse (WWC) does not give a single dropout-prevention program its top rating of "strong effects" (defined as several randomized experiments or quasi-experiments all pointing in the same direction, or one large randomized experiment). The Coalition for Evidence-Based Policy does not list a single program for addressing high school graduation rates among its "Top Tier" of programs.

Our understanding of how to reduce youth violence is little better. The influential Blueprints for Violence Prevention reviewed over 900 studies; the total number of "model programs" found to reduce criminal involvement among teens was just *four*. Three of these model programs work with youth already in the criminal justice system and are relatively costly (Multi-Systemic Therapy, which costs $4,500 per participant, Multi-Dimensional Treatment Foster Care, $27,300 per youth, and Family Functional Therapy, FFT, which cost $1,600-$5,000 per youth) – and so may not be scalable. And the empirical evidence is somewhat less

compelling than one might have imagined given their designations as "model programs" for two of these four model programs: FFT and Big Brothers / Big Sisters (BB/BS).[40]

[40] While a meta-analysis suggests that FFT is a cost-effective crime-reduction strategy (Drake, Aos & Miller 2009), some of the relevant studies did not involve random assignment (Barnoski & Aos 2004; Barton, et al. 1985; Gordon, Graves & Arbuthnot 1995) and others use sample sizes under 100 (Alexander & Parsons 1973; Klein, Alexander & Parsons 1977). The largest randomized study (917 families) found no differences overall between treatment and control youth; reduced recidivism only occurred for the treatment youth who happened to receive high-fidelity versions of the program (Sexton & Turner 2010). Yet youth were not randomly assigned to the high-fidelity group, so we cannot be certain their reduced recidivism was due to the program (particularly given that this group had significantly different criminal histories at baseline). The study of BB/BS relies on youth self-reports to measure outcomes, and so may confound program effects on behavior with effects on self-reporting bias if program youth are worried about disappointing their mentors (Spelman 1994).

Appendix B:
Selection of study schools and students

Our team originally recruited 16 CPS schools to participate in the study. Four of those schools run separate achievement academies within the same building. These are large and distinct schools for students facing academic or social barriers to conventional academic advancement. Since these achievement academies ran distinct treatment groups and we randomized them separately, we treat them as separate schools – and so essentially began the study with 20 schools total. The program was never actually implemented in one school, and one school was excluded for problems with randomization. (We did not have a full set of baseline characteristics available at the time we had to carry out randomization in this school, and so randomized 83 youth to treatment and control without being able to construct the risk index described below. *Ex post* analysis indicated there were statistically significant imbalances between the treatment and control groups on baseline characteristics). Our main sample therefore consists of 18 schools.

Five of these schools are elementary schools, which in Chicago serve students in grades K-8, while 13 are high schools serving grades 9-12. Sample schools were among the lowest-performing in CPS; seven have average GPAs of 2.25 or below (Roderick, Nagaoka & Allensworth 2006). Three are currently "turnaround" schools, chosen for major reform due to consistently low student performance.

To construct our study sample frame, CPS provided us with administrative data on all male students they expected to attend the study schools and be enrolled in grades 7-10 during the 2009-10 academic year (AY), a total of 3,669 students. We focused on males because of their very disproportionate involvement in serious inter-personal violence in Chicago (as in every other U.S. city). Based on discussions with the non-profit organizations running the intervention, prior to randomization we excluded some students according to their baseline (AY 2008-9) characteristics:[41]

1. Youth who seem to have stopped going to school, and so were unlikely to attend school frequently enough to benefit from a school-based program. A total of 268 students (about 5 percent of our initial sample frame) were excluded because they missed more than 60 percent of days during AY 2008-9 and received a grade of "F" in at least 75 percent of their courses.

2. Students with specific Individualized Education Program (IEP) designations for serious discrete conditions including autism, emotional and behavioral disorders, speech and language disabilities, "educatable mentally handicapped," traumatic brain injury, and diagnosed emotional and behavioral disorders. Service providers

[41] Staff of the non-profit organization that ran the intervention determined after the initial randomization that they could not effectively serve youth who were significantly older than grade level—which staff operationalized as all youth born before October 1, 1992. A total of 153 youth were therefore "never takers" because the provider decided not to offer them treatment, although were still kept in the study sample to preserve randomization. At some locations, this resulted in suboptimal sample sizes for the number of youth receiving treatment. A total of 152 age-eligible control youths from 10 schools were then selected at random to replace these age-ineligible youth, though we treat them as controls for the analysis to maintain the original randomization.

determined that youth with these specific diagnoses were unlikely to benefit from the BAM curriculum. A total of 294 youth (just over 5 percent) were excluded for this reason. Less intensive IEP designations were not used as a study exclusion criterion. Indeed, roughly 20 percent of our final study sample had some sort of IEP designation, most commonly for the general category of learning disability.

We then ranked all the remaining students in our target CPS schools on the basis of a risk index, which was a single-factor composite of whether a student was at least one year older than his assigned grade level, the number of classes for which a student had received a grade of "F" during AY 2008-9, the number of unexcused absences during AY 2008-9, and the number of in-school suspensions during AY 2008-9. A large number of students were missing academic achievement test scores for AY 2008-9, due to some combination of CPS testing schedules (not all grades are subject to standardized testing each year) and student absences during testing days. Due to the large number of missing items for this variable, test scores were not used selecting the study sample.

We then calculated the number of students needed in each school for the study sample (treatment and control), selected that many students in descending order on our risk index, and randomized those selected students to one of four conditions (in-school only, after-school only, both, or neither) within school (a block-randomized design with schools as blocks).

Appendix C:
Additional details on administrative data sources

In this section, we discuss in more detail the administrative data sources we use to measure outcomes in both domains.

Our main schooling outcomes come from longitudinal student-level records obtained from CPS for the program year (AY 2009-10) and follow-up year (AY 2010-11), with controls in our regression for student outcomes during the pre-program year (AY 2008-9). Because our original sampling frame was drawn from CPS data, we could use students' CPS identification numbers to directly access their data rather than using probabilistic matching. We used these data to form a summary index of schooling outcomes, equal to an (unweighted) average of days present, GPA, and persistence in school (enrollment status at the end of the 2009-10 academic year[42]), each normalized to Z-score form using the control group's distribution. Our index of schooling outcomes does not include standardized test scores. CPS by design does not administer standardized tests to all grades (particularly older grades). Thus, more than half of all students in our study sample are missing test scores (similar shares for treatment and control groups). Our index also does not include administrative records on school disciplinary actions; we have heard conflicting accounts from multiple sources familiar with CPS records about whether disciplinary actions are inconsistently reported and recorded.[43]

To measure criminal behavior by program participants, we use electronic arrest records (or "rap sheets") from Illinois State Police (ISP), obtained through the Illinois Criminal Justice Information Authority (ICJIA) for research purposes. The ISP maintains a database of Illinois arrests, collected from local police departments which are required by law to report all juvenile felony arrests, and optionally class A and B misdemeanors. Some local police departments also voluntarily submit information that is not subject to mandatory reporting requirements. Police departments in Illinois use biometric (fingerprint) identification systems at the time of arrest. This means that even if a youth lies about his identity when arrested, as long as his fingerprint is on record and he has ever reported his actual name and date of birth at any point for any of his arrests, then an arrest in which he misreports his individual identifying information will still show up in our data attached to the correct state identification number.

The ISP arrest records are intended to capture arrests of people below the age of majority within the criminal justice system (juvenile arrests), as well as to those who are above the age of majority (age 17 and above for those charged with a felony; starting in 2010, 17-year-olds charged with a misdemeanor offense are considered to be a juvenile). Over the course of the 2000s, the ISP data system's coverage of juvenile arrests reportedly improved a great deal. A comparison of "age-crime curves" (arrest rates by age) for Chicago youth compared to urban

[42] We operationalize end-of-year enrollment as either having at least one grade on record for the second semester of the year, or as being marked as transferring to a non-CPS school. We could alternatively use official CPS enrollment status, which although subject to incentives to misreport, does not appreciably change the results.

[43] Using data on all students who were present for at least half a day in CPS (thus having a chance to commit a disciplinary infraction), three of four disciplinary measures have negative point estimates (total number of infractions, number of serious infractions, and number of out of school suspensions) ranging from -0.01 to -0.04 standard deviations; in-school suspension has a positive point estimate of 0.03. None of these effects is statistically significant.

youth in other cities where the age of majority is different suggests no unusually large "jump" in arrest rates for Chicago youth at the age at which teenage arrestees are automatically sent to the adult justice system (Kling, Ludwig & Katz 2005).

ICJIA used a probabilistic-matching software (Merge Tool Box) to match the list of study names and dates of birth to their Criminal History Record Information database. Soundex algorithms were used to help match first and last names across error-prone administrative databases. We use the arrest histories for any individual with a match quality score above 32.29501, or whose identifying information matched on at least 4 of the 5 available fields (first name, last name, day of birth, month of birth, and year of birth). Results using a lower quality threshold (31.63083) that incorporate another 167 matches are less precise but qualitatively similar (see Table A2).

Because previous studies often find more pronounced impacts of policy interventions on violent crimes (particularly impulsive crimes such as assault) than on other crimes (Deming 2011; Evans & Owens 2007; Kling, Ludwig & Katz 2005; Lochner & Moretti 2004; Weiner, Lutz & Ludwig 2009)) and because associated social harms are so varied across crime types, we examine arrests separately for different offense categories. For each arrest incident, we select the most severe charge associated with the incident. In most cases this is a charge recorded at the time of arrest, although occasionally the State's Attorney files a charge more severe than those originally recorded at the police station. We classify crimes as violent, property, drug, and other, as follows:

(1) *Violent crimes* include murder, rape, assault, robbery, threats/harassment, and kidnapping.
(2) *Property crimes* include larceny, burglary, and auto theft.
(3) *Drug crimes* include possession or dealing charges.
(4) *Other crimes* include trespassing, fencing, bribery, animal cruelty, weapons violations, DUIs, disobeying or avoiding law enforcement officers, disorderly conduct, arson, prostitution, criminal neglect, parole violations, underage or public drinking, vandalism, and miscellaneous offenses.

We exclude motor vehicle crimes, including driving with a suspended license, reckless driving, and other driving/traffic related offenses, from our analysis. These are rare in our data (due largely to the ages of those in our sample and presumably also to the lack of cars among so many of the low-income families from which these youth come).

Appendix D:
Methods for addressing missing outcome data in CPS schooling records

While the proportion of treatment and control youth missing data during the post-randomization years is statistically indistinguishable,[44] the amount of missing data increases over the course of the post-program period. Because we selected students using AY 2008-9 CPS data, there is basically no missing data for that pre-program year. By construction, no one has missing values on our measure of school persistence for any of the post-randomization years, and so all observations have at least one non-missing element of our outcome index. For the program year itself (AY 2009-10), 274 out of 2,740 students are missing GPA information (10 percent); 80 of those 274 are also missing attendance information. In the post-program year (AY 2010-11), we are missing GPA information for 903 of 2,740 students (33 percent); of that group, we are missing data on days present as well as GPA for 431 students. [45]

The results presented in our main tables assume that data are missing completely at random (MCAR), that is, the likelihood of missingness is unrelated to both observable and unobservable attributes of youth, including potential outcomes. Our main results use the approach from Kling, Liebman, and Katz (2007) that assigns the group average (treatment or control) to missing values on those elements of the index, which is equivalent to averaging the effects of running separate regressions on each element of the outcome index using just those observations with non-missing observations on the index, but with added power.

In Table V of the main text, we show the results of a lower-power MCAR approach: limiting the regression to complete cases (listwise deletion), which throws away information for cases that have only some missing data. We also re-calculate our estimates using different imputation approaches that relax the MCAR assumption and rely instead on the assumption that missingness is missing at random (MAR) – that is, for reasons that may be related to observable youth characteristics but not to unobservables. We first use inverse probability weighting (IPW) to weight the complete-case sample so that the distribution of baseline characteristics in that sample matches the distribution of baseline characteristics in the full sample. To calculate these weights, we regress an indicator for having complete data on a parsimonious set of baseline covariates (specifically, GPA, being over age 17, and dummy variables for being black or having been arrested). We then use the inverse of that predicted probability as a weight in our outcome regressions, which include all baseline covariates. As long as this prediction equation is correctly specified, the IPW approach provides consistent estimates of the treatment effect under MAR.

It is worth noting that, as is often the case with IPW (see, e.g., discussions in Cox 1991; Puma, et al. 2009), specifying the prediction equation involves a tradeoff between bias and

[44] When we regress an indicator for whether or not the student has missing data against an indicator for treatment assignment, controlling for school fixed effects, the coefficient on treatment assignment for missing data in the program year (AY 2009-10) equals β=0.011 (standard error 0.12), p=0.349, and for the post-program year of AY 2010-11, equals β=0.009 (0.018), p=0.608.

[45] Our CPS data include some information about the why students stop showing up in the CPS data system in the form of "leave codes," which are intended to capture the reason why a student left the district. Yet the reliability of these type of leave codes is open to question in administrative schooling data given schools' possible incentives to strategically misreport. In most districts, schools receive funds based on fall enrollments and are held accountable for the share of students who are officially counted as dropping out.

variance. When we use too many additional baseline covariates that appear to be related to the probability of missingness, observations with very low predicted-response probabilities get implausibly large and unstable weights. This makes both the point estimate and the standard errors sensitive to the choice of variables in the prediction equation. While it seems likely that the prediction equation we use here does not fully satisfy MAR (that is, both observed and unobserved characteristics that affect missingness are not included in the prediction equation), we exchange this problem for increased stability, turning to other imputation approaches to ensure that our results are robust.

One simple imputation approach we use is to assign zeros for attendance and grades for students who are recorded as not being enrolled in the CPS system at the end of the academic year – what Puma et al. (2009) call "logical imputation" (or "deductive imputation"). A drawback of this very simple logical-imputation approach is that some youth who are not enrolled in the CPS system at the end of the academic year will not have dropped out (for example, they may have transferred to a private school in Chicago, or to a suburban public school, or to a Chicago public school that did not record their new enrollment), and so their attendance and grade information is truly missing rather than zero. A slightly more sophisticated logical imputation approach is to make use of the "leave codes" in the CPS data that record information about the reasons why students are no longer in the system. One potential drawback to these records is they themselves may not be perfectly accurate, in part because schools may have incentives to over-state enrollment figures and under-state dropouts.

A third imputation approach we employ is multiple imputation (MI), which incorporates all the uncertainty involved in making imputations. MI uses a Bayesian approach, drawing from the conditional predictive distribution of the missing variable(s) to impute values of the missing data, and iterating the process to analyze *m* of these simulated data sets (Little & Rubin 2002; Puma, et al. 2009). To see the intuition, let us first consider the uncertainty involved in a single imputation. Imagine that we impute missing values by regressing each of the variables (for observations that are non-missing) against all the other variables, then using the predicted value from those regression parameters as the imputation for missing observations. This prediction would lie exactly on the regression line and so would overstate our certainty about its value. To adjust for this uncertainty, we could also add to the predicted value a randomly selected residual from the regression, so that the variation in the imputed values is the same as with the observed values. But there is the additional uncertainty involved with the imputation; if we repeated this process, we would get a somewhat different imputed data set. To account for the uncertainty associated with using imputed values, this imputation procedure is repeated numerous times, and separate impact estimates are calculated using each dataset created by the multiple iterations. The average of the estimates for the desired statistic is the MI estimate, and the standard error accounts for both within- and between-imputation variances.

There is, however, once source of uncertainty remaining: we have used a regression to predict the missing data, but the coefficients in that regression are estimates, not true parameters. MI takes a Bayesian approach to incorporating this uncertainty. Using some starting value for the parameters of the imputation regression (the betas and sigma-squared), we predict one set of missing data to form a complete imputed data set as just described. We then calculate new estimates of the betas and sigma-squared by re-running the regression on the imputed data, after

which we update the distribution of both parameters using these new parameter estimates. Next we make new draws of the betas and sigma-squared from the updated distributions, re-run the prediction regression, and so forth. MI iterates this process, forming a usable imputed data set after some set number of iterations, then repeating the process to form the *m* simulated data sets. Under MAR, MI provides consistent estimates that both make use of all the data to predict missing values and incorporate the uncertainty involved in the imputation.

We present results using MI with $m = 10$ imputed data sets. We run separate imputations for the treatment and control groups for missing GPA and attendance information, then recalculate the academic engagement index within each data set and re-run our regressions.[46] Imputing outcomes separately for treatment and control groups avoids injecting correlation with the treatment indicator into the imputation (Puma, et al. 2009). There is also evidence that the baseline characteristics predict missingness differently for the treatment and control groups, at least in the follow-up year. We cannot reject the null that the relationship between baseline characteristics and missingness is the same across treatment status during the program year (regressing a missing indicator on baseline characteristics and the interaction of those covariates with treatment results in a joint F-test on the treatment interactions of 1.24, p=0.1585). However, the interaction of baseline characteristics with treatment status does seem to matter during the follow-up year (F(=1.55, p=0.0225)). Since baseline characteristics may have differing effects on missingness by group, it is sensible to allow their effects on the imputed values to vary by group as well. These estimates account for both within- and between- imputation variances. We show in Table IV that the results from MI are similar to those using the Kling, Liebman and Katz approach, so we present the latter in our main tables for simplicity.

It is also possible that the data are missing for reasons related to unobservable as well as observable youth attributes – that is, the data are not missing at random (NMAR). Given this possibility, we also present bounds that use Lee's trimming approach (2009). The intuition is that although we cannot identify which specific treatment-group youth are observed in the dataset only because of treatment (e.g., are selected into the sample), we can assume the highest or lowest observed values of the outcomes belong to the group selected into the observed-data sample because of treatment assignment, exclude those values, and calculate the treatment-control difference in each case.

Specifically, suppose that a certain proportion (p) of the treatment group is selected into the sample because of treatment but would have missing outcome data if assigned to the control condition. The mean treatment outcome we actually observe is a weighted average of this group and those 1-p treatment youth who would always be observed regardless of random assignment. To be comparable to the unselected control group, we would like to remove these p selected treatment observations. But we do not observe which p members of the treatment group are selected into the sample. We can, however, create worst-case scenarios where the highest (or lowest) p values of the outcome variable in the treatment group belong to these sample-selected cases. Trimming off the highest (lowest) p^{th} quantile of outcome values (call them Y) creates a lower (upper) bound for the true value of the non-sample-selected cases, because the mean Y for any possible subset of the treatment group of size 1-p can not be smaller (larger) than the mean Y

[46] Specifically, we use chained predictive mean matching for days present and GPA since both are truncated continuous variables (neither can be less than zero).

of the smallest 1-p values. Provided that treatment assignment only influences sample selection in one direction – the monotonicity assumption – this procedure provides one way to bound the treatment effect under a fairly extreme sample selection mechanism. The monotonicity assumption underlies every selection model that is based on a latent-variable threshold for participation (Lee 2009), and means treatment can only encourage *or* discourage selection into sample, but not both.

The Lee bounds make fairly extreme assumptions about the nature of the data missingness process. For example the lower bound is created based on the extreme assumption that academic outcomes are perfectly negatively correlated with the latent propensity to remain in the sample (see Lee 2009 for discussion). Since some proportion of the unobserved students transferred to another school rather than dropped out (the noisy CPS leave codes suggest about half), this assumption is quite extreme. There are also reasons to question the method's monotonicity assumption in this case – namely that a higher fraction of treatment youth are in the sample overall and in most of our schools, but not in every school we study. Overall, we observe a higher proportion of treatment than control youth in our CPS outcome data, suggesting treatment encourages selection into the sample. However, if we break down the sample by school, it is not clear that the selection mechanism consistently works this way. In the program year, four of the 18 schools have higher proportions of control youth observed, and three have no difference between treatment and controls. In the follow-up year, fully half of the schools have more observed control than treatment youth, while the other one-half have more treatment youth. The fact that treatment assignment has no consistent relationship with sample selection across all of our schools might diminish the concern that treatment status is systematically correlated with the missingness mechanism (as well as call into question the monotonicity assumption underlying the Lee approach).

Appendix E:
Methods and results for testing for treatment-effect heterogeneity across schools

In a regression that interacts all school dummies with a treatment indicator, we cannot reject the null that the treatment effects at each school are equal (for the program-year academic ITT with covariates, F = 0.56, p = 0.9238; for the violent crime arrests with covariates, F = 1.18, p = 0.2708).

Non-economists tend to prefer to perform this test using hierarchical linear modeling, a form of a random coefficients model, in part for the efficiency gain that comes with precision-weighting the blocks (schools). In this case, the model would be:

$$(6) \quad Y_{is} = \beta_0 + \beta_1(Z_{is} - \overline{Z}_{i.}) + r_{is}$$

where

$$\beta_{0s} = \gamma_{00} + u_{0s}$$

$$\beta_{1s} = \gamma_{10} + u_{1s}$$

and where Y_{is} is the outcome variable during the program year, $Z_{is} - \overline{Z}_{i.}$ is a school-demeaned (group-centered) treatment indicator from the baseline year, and r_{is} is the error term.[47] For simplicity we do not include covariates, and we let each school have a random intercept and treatment effect.

This empirical strategy produces very similar substantive results as OLS, with the ITT treatment effect on the program year academic composite equal to 0.058 (p = .04) and equal to 0.075 (p = 0.01) for the follow-up year academic composite. Using an over-dispersed Poisson model for violent crime, we find an ITT effect of -0.217 (p = 0.08) for violent crime arrests, very similar to the QMLE Poisson estimate of -0.201 reported in the main text. But here we can directly test whether the variance of the slopes – i.e., the treatment effect's variance – across schools is zero.

We find that the treatment effect's variance is quite small: 0.00005 for the program year academic treatment effect, 0.0002 for the follow-up year academic treatment effect, and 0.0003 for the violent crime arrest effect. We fail to reject the null hypothesis that the variance of each of the treatment effects is zero. For the violent crime arrests, the chi-square test for whether the variance of treatment effects across schools is zero equals 11.62 (p > 0.5). The result is similar for the academic index (chi-square = 9.59, p>0.5 for the program year and chi-square = 6.50, p>0.5 for the follow-up year).

Given that there are only 18 schools in the sample, this should not be interpreted as strong evidence for constant treatment effects; our power to detect differential effects is limited, and our schools are more similar to each other than schools in other cities may be to our sample. The results do mean, however, that the school-level differences in observed treatment effects were not large enough to statistically differentiate.

[47] Demeaning the independent variables provides the same adjustment for blocking to the covariates as including school fixed effects.

Appendix F.
Benefit-Cost Estimates

We calculate the benefits of the program in four parts: the benefits from the realized crime reduction during the program year (both direct savings to the criminal justice system and the broader social savings) and the benefits from the predicted increase in graduation based on achievement increases (benefits to the government from increased revenues and decreased social service use, as well as earnings benefits to the participant).

We know of no national estimates for the cost to the criminal justice system from processing an average arrest, which includes police processing as well as the costs of later stages of prosecution for some subset of arrested youth (detention, incarceration, probation, etc.). We construct these estimates ourselves from a range of sources, using Chicago-specific data on average costs and the probability of incurring each cost when possible, and relying on other city's estimates (mostly New York City) when Chicago data are unavailable (Hughes & Bostwick 2011; Illinois Juvenile Justice Comission 2011; New York City Independent Budget Office 2008). We find that the cost of an average arrest to the criminal justice system is between $5,770 and $6,524. This is likely a conservative estimate due to the exclusion of court and policing costs; a similar calculation for North Carolina found each arrest costs an average of $7,300 (Governor's Crime Comission 2009). The main LATE estimate of the treatment effect for total number of arrests is -0.1865 (p=0.087). This implies the program saved 18.65 percent of the cost of one arrest per participant, or between $1,076 and $1,217 in direct criminal justice costs.

While these costs are important, especially to policymakers, it is clear that the total benefits to society from reductions in crime are far broader than just the tangible costs averted to the criminal justice system. Indeed, previous research on gun crimes in particular suggests that intangible benefits from reduced crime may far outweigh tangible benefits (Cook & Ludwig 2000). Monetizing the value to society from reduced crime is challenging in part because of complicated conceptual questions like the costs to society from drug use by individuals. Other practical problems arise because the costs of crime are so skewed, with homicide being at least an order of magnitude more costly than any other crime in most studies, and because there is no perfect way to monetize the intangible costs of crime.

Conceptually, the ideal way to measure the value to society from reductions in crime in the future is from an *ex ante* perspective – what is the aggregate sum of the public's willingness to pay (WTP) for reducing the risk of crime victimization in the future? Contingent valuation (CV) surveys in principle are capable of capturing this WTP value, but in practice many people are understandably nervous about relying on survey responses to hypothetical questions about what people would be willing to pay to reduce crime. An alternative approach has been to rely on jury award data. But jury awards adopt an ex post perspective after a victim is identified and so are problematic from a conceptual perspective, and a very small and unusual subset of criminal events result in civil litigation for damages (see Cohen 2005; Cook & Ludwig 2000).

We start by following the basic approach from Kling, Ludwig, and Katz (2005), assigning each type of crime the social costs estimated by Miller, Cohen and Wiersema's (1996) that rely on jury award data, and examining the sensitivity of our estimates to how we handle the

social costs of homicides. We note that since our estimates are based on arrests, and not all criminal offenses result in arrest, our estimates may understate the total social benefits from averted crimes among treatment youth (although it is also true that not all arrested youth actually committed the crime for which they were arrested). Because different types of crimes will impose different costs on society, we assign each crime category a unique cost, aggregate the costs across types of arrests for each youth, and use that individual cost-of-crime total per youth as the outcome variable.

The top panel of Appendix Table A5 suggests that participation in the intervention reduces the social costs of crime per youth between $7,140 and $11,983 per youth, or 70 to 80 percent of the control complier mean. The estimates are very imprecise largely because of the very high costs of homicide and how few homicide arrests there are in the data. Although these estimates are not quite significant at the 10 percent level, their magnitudes are noteworthy; a $1,100 per youth program that has an 85 percent likelihood of saving from $7-12,000 per participant is still likely to be of some interest to public policymakers (see for example Cook & Ludwig 2006).

As the bottom panel of Appendix Table A5 shows, using estimates for the costs of crime from contingent valuation studies (Cohen, et al. 2004), which tend to be higher than those from jury awards, yields even larger estimates for the savings from our program due to reductions in youth crime. Although these CV estimates for the social costs of crime are somewhat controversial, it is worth noting that using this alternate calculation for violent crimes only significantly increases our estimates of social savings due to the intervention. Since these estimates are based on people's willingness to pay to avoid violent crimes, they in theory already incorporate how much people are willing to pay for reductions in future crime that is driven by their desire to avoid having to pay in taxes in the future for direct criminal justice system costs. The benefit calculations in the high-end estimate that use these figures (main text Table IX) therefore subtract the direct taxpayer benefits from the total in Table A4.

Table IX in the main text also shows a lower-bound estimate for crime-reduction benefits. This uses the LATE estimate on total arrests during the program year from the most conservative bounding exercise, which adjusts for possible under-reporting of after-school attendance (β = -0.1206 (0.0704), p = 0.087), and the lowest cost of crime estimates from Miller, Cohen, and Wiersma (1996) (trimming murder costs by half).

Since the students in our sample are too young to have graduated, and most estimates of the benefits of education are based on graduation, calculating the potential benefits of increased school performance is more difficult. The first challenge is to extrapolate the implied effect on future graduation rates from the estimated impacts on schooling outcomes that we observe during our two-year study period. We find no significant difference in school persistence to date, but our measures of persistence are noisy and many students are still too young to officially drop out of school. We do, however, observe significant changes in GPA and course failures (and not quite significant changes in days present). To our knowledge, there are no causal estimates of how changes in GPA or days present for grades 7 – 10 affect later school completion.

The Consortium report discussed in the main text provides correlational estimates of the relationship between graduation rates and GPA from longitudinal data. We use these rates to estimate our sample's probability of graduating based on their GPA, using the same KLK approach to impute missing GPAs as for our main index[48]. We then use predicted graduation rate as a dependent variable. The size of the LATE improvements on predicted graduation rates (+0.0521 for the program year and + 0.1010 for the following year) imply 10 and 22 percentage point increases respectively, compared to control complier means of 53 and 45 percent. Using the lower-bound LATE estimates instead (+0.0336 and +0.0651 for the two years) translates to a graduation increase of 6.6 and 13.6 percent by year, from the relevant control-complier baselines of 50 and 48 percent respectively. These estimates provide our range of 7 to 22 percent increased graduation reported in the main text. Because we think the more recent data from the follow-up year might be a better indication of future graduation, we use the estimates from the 10-11AY in monetizing the potential graduation benefits.

We would emphasize that this prediction involves a great deal of extrapolation (it is based on correlational estimates for the entire district of 9th graders rather than a causal estimate for our particular population of 7th – 10th grade boys). Our intent is to provide a sense of the potential magnitude of the changes in the academic index, not to put a great deal of stock in the specific estimates themselves.

Our second step is to then estimate the total monetized benefits to society from increased high school graduation rates. We rely on estimates for the lifetime benefits of high school graduation from Levin et al. (2007), who calculate the present discounted value of each graduate's earnings relative to a dropout's, and how much a graduate contributes over his lifetime in terms of additional taxes and lower health-care and welfare costs. Given the characteristics of our study sample, we use Levin et al.'s estimates for black and Hispanic males and inflate the figures to 2010 dollars. These calculations results in the figures in the bottom panel of Table IX.

[48] Note that this likely understates the amount of dropout, since some youth may be missing GPAs because they have already dropped out. This possibility is consistent with the fact that the average graduation rate for everyone in the schools we study (boys and girls) is only 40 percent; it seems unlikely that our sample of boys would have higher than average expected graduation rates. However, results that instead assign a 0 probability of graduating to youth who are missing GPA data and who CPS reports are no longer enrolled are very similar to those reported here.

REFERENCES

Alexander, J.F., and B.V. Parsons, "Short-term behavioral intervention with delinquent families: Impact on family process and recidivism," Journal of Abnormal Psychology, 81 (1973), 219.

Allensworth, E.M., and J.Q. Easton, *What Matters for Staying On-Track and Graduating in Chicago Public High Schools* (Chicago: Consortium on Chicago School Research, University of Chicago, 2007).

Allensworth, Elaine, "Update to From High School to the Future: A first look at Chicago Public School graduates' college enrollment, college preparation, and graduation from four-year colleges," (Chicago, IL: Consortium on Chicago School Research, 2006).

Anderson, ML, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," Journal of the American Statistical Association, 103 (2008), 1481-1495.

Angrist, J.D., G.W. Imbens, and D.B. Rubin, "Identification of causal effects using instrumental variables," Journal of the American Statistical Association, 91 (1996), 444-455.

Antoni, Michael H., Stacy Cruess, G. Dean, Mahendra Kumar, Susan Lutgendorf, Gail Ironson, Elizabeth Dettmer, Jessie Williams, Nancy Klimas, Mary Ann Fletcher, and Neil Scheidermann, "Cognitive-behavioral stress management reduces distress and 24-hour urinary free cortisol output among symptomatic HIV-infected gay men," Annals of Behavioral Medicine, 22 (2000), 1532-4796.

Aos, S., M. Miller, and E.K. Drake, "Evidence-Based Public Policy Options to Reduce Future Prison Construction, Criminal Justice Costs, and Crime Rates," (Olympia: Washington State Institute for Public Policy, 2006).

Armstrong, T.A., "The effect of moral reconation therapy on the recidivism of youthful offenders," Criminal Justice and Behavior, 30 (2003), 668.

Barnoski, R.P., and S. Aos, "Outcome evaluation of Washington State's research-based programs for juvenile offenders," (Washington State Institute for Public Policy, 2004).

Barrett, P. M., A. L. Duffy, M. R. Dadds, and R. M. Rapee, "Cognitive-behavioral treatment of anxiety disorders in children: Long-term (6-year) follow-up," Journal of Consulting and Clinical Psychology, 69 (2001), 135-141.

Barton, C., J.F. Alexander, H. Waldron, C.W. Turner, and J. Warburton, "Generalizing treatment effects of functional family therapy: Three replications," The American Journal of Family Therapy, 13 (1985), 16-26.

Beck, J.S., *Cognitive therapy: Basics and beyond* (The Guilford Press, 2011).

Benjamini, Y., and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," Journal of the Royal Statistical Society. Series B (Methodological), (1995), 289-300.

Benjamini, Y., A.M. Krieger, and D. Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," Biometrika, 93 (2006), 491-507.

Birmaher, B., D. A. Brent, D. Kolko, M. Baugher, J. Bridge, D. Holder, S. Iyengar, and R. E. Ulloa, "Clinical outcome after short-term psychotherapy for adolescents with major depressive disorder," Archives of General Psychiatry, 57 (2000), 29-36.

Bloom, H.S., L.L. Orr, S.H. Bell, G. Cave, F. Doolittle, W. Lin, and J.M. Bos, "The benefits and costs of JTPA Title II-A programs: Key findings from the National Job Training Partnership Act study," Journal of Human Resources, (1997), 549-576.

Bloom, Howard S., "Accounting for No-shows in Experimental Evaluation Designs," Evaluation review, 8 (1984), 225-246.
Borghans, L, AL Duckworth, JJ Heckman, and B Ter Weel, "The Economics and Psychology of Cognitive and Non-Cognitive Traits," Journal of Human Resources, (2007).
Bowles, S, H Gintis, and M Osborne, "The determinants of earnings: A behavioral approach," Journal of Economic Literature, 39 (2001), 1137-1176.
Brent, D. A., D. Holder, and D. Kolko, "A clinical psychotherapy trial for adolescent depression comparing cognitive, family, and supportive treatments," Archives of General Psychiatry, 54 (1997), 877-885.
Cameron, A.C., J.B. Gelbach, and D.L. Miller, "Bootstrap-based improvements for inference with clustered errors," The Review of Economics and Statistics, 90 (2008), 414-427.
Campbell, FA, CT Ramey, E Pungello, J Sparling, and S Miller-Johnson, "Early childhood education: Young adult outcomes from the Abecedarian Project," Applied Developmental Science, 6 (2002), 42-57.
Survey Methodology (http://help.ccsrsurvey.uchicago.edu/customer/portal/articles/94362-survey-methodology,
Clampet-Lundquist, Susan, Stefanie DeLuca, and Kathryn Edin, "Title," Johns Hopkins University Working Paper, 2012.
Clarke, G., H. Hops, and P. M. Lewinsohn, "Cognitive-behavioral group treatment of adolescent depression: prediction of outcome," Behavioral Therapy, 23 (1992), 341-354.
Cohen, M.A., *The costs of crime and justice* (Psychology Press, 2005).
Cohen, Mark, Roland Rust, Sara Steen, and Simon Tidd, "Willingness to pay for crime control programs.," Criminology, 42 (2004), 86-106.
Conduct Problems Prevention Research Group, "The Effects of the Fast Track Preventive Intervention on the Development of Conduct Disorder Across Children," Child Development, 82 (2011), 331-345.
Cook, P, and J Ludwig, *Gun Violence: The Real Costs.* (New York: Oxford University Press, 2000).
Cook, P. J., and J. Ludwig, "The social costs of gun ownership," Journal of Public Economics, 90 (2006), 379-391.
Cox, B., "Weighting Survey Data for Analysis," Presentation for the ASA continuing education program, (1991).
CPD, "Annual Report," (Chicago, 2011a).
---, "Chicago Murder Analysis," (Chicago, 2011b).
Cunha, F., and J. Heckman, "The Technology of Skill Formation," American Economic Review, 97 (2007), 31-47.
Currie, Janet, and Duncan Thomas, "Does Head Start Make a Difference?," American Economic Review, 85 (1995), 341-364.
Deming, D, "Early childhood intervention and life-cycle skill development," American Economic Journal: Applied Economics, 1 (2009), 111-134.
Deming, David, "Better Schools, Less Crime?," Quarterly Journal of Economics, 126 (2011), 2063-2115.
Dodge, K.A., "Do social information-processing patterns mediate aggressive behavior?," (2003).
Dodge, Kenneth A. , John E. Bates, and Gregory S. Pettit, "Mechanisms in the cycle of violence," Science, 250 (1990), 1678-1683.

Drake, E.K., S. Aos, and M.G. Miller, "Evidence-based public policy options to reduce crime and criminal justice costs: implications in Washington State," Victims and Offenders, 4 (2009), 186.
Durlak, J.A., R.P. Weissberg, A.B. Dymnicki, R.D. Taylor, and K.B. Schellinger, "The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions," Child Development, 82 (2011), 405-432.
Dynarski, M., P. Gleason, A. Rangarajan, and R.G. Wood, *Impacts of dropout prevention programs: Final report* (Mathematica Policy Research, Incorporated, 1998).
Evans, W.N., and E.G. Owens, "COPS and Crime," Journal of Public Economics, 91 (2007), 181-201.
Farrell, AD, AL Meyer, TN Sullivan, and EM Kung, "Evaluation of the Responding in Peaceful and Positive Ways (RIPP) seventh grade violence prevention curriculum," Journal of Child and Family Studies, 12 (2003), 101-120.
Farrell, AD, AL Meyer, and KS White, "Evaluation of Responding in Peaceful and Positive Ways (RIPP): A school-based prevention program for reducing violence among urban adolescents," Journal of Clinical Child & Adolescent Psychology, 30 (2001), 451-463.
Gaab, J., N. Blattler, T. Menzi, B. Pabst, S. Stoyer, and U. Ehlert, "Randomized controlled evaluation for the effects of cognitive-behavioral stress management on cortisol responses to acute stress in healthy subjects," Psychoneuroendocrinology, 29 (2003), 767-779.
Garbarino, J, "Lost boys: Why our sons turn to violence and how to save them," (New York: Free Press, 1999).
Garces, Eliana, Duncan Thomas, and Janet Currie, "Longer-term effects of Head Start," American Economic Review, 92 (2002), 999-1012.
Goldin, Claudia, and Lawrence F. Katz, *The Race between Education and Technolog* (Cambridge, MA: Belknap Press of Harvard University Press, 2008).
Gordon, D.A., K. Graves, and J. Arbuthnot, "The effect of functional family therapy for delinquents on adult criminal behavior," Criminal Justice and Behavior, 22 (1995), 60-73.
Governor's Crime Comission, "Juvenile Age Study: A study of the impact of the jurisdiction of the Department of Juvenile Justice and Delinquency Prevention," (Final Report to the Governor of North Carolina, 2009).
Greenwood, P., "Prevention and intervention programs for juvenile offenders," The future of Children, 18 (2008), 185-210.
Grossman, JB, and JP Tierney, "Does mentoring work?: An impact study of the Big Brothers Big Sisters program," Evaluation review, 22 (1998), 403.
Gundersen, Knut, and Frode Svartdal, "Aggression replacement training in Norway: Outcome evaluation of 11 Norwegian student projects," Scandinavian journal of educational research, 50 (2006), 63-81.
Harrington, Nancy G, Steven M Giles, Rick H Hoyle, Greg J Feeney, and Stephen C Yungbluth, "Evaluation of the All Stars character education and problem behavior prevention program: Effects on mediator and outcome variables for middle school students," Health Education & Behavior, 28 (2001), 533-546.
Heckman, James J., and Paul A. LaFontaine, "The American High School Graduation Rate: Trends and Levels," Review of Economics and Statistics, 92 (2010), 244-262.
Heckman, James J., and Yona Rubinstein, "The Importance of Noncognitive Skills: Lessons From the GED Testing Program," American Economic Review, 91 (2001), 145-149.

Heckman, James J., Jora Stixrud, and Sergio Urzua, "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," Journal of Labor Economics, 24 (2006), 411-482.
Heckman, JJ, L Malofeeva, R Pinto, and PA Savelyev, "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes," Unpublished manuscript, University of Chicago, Department of Economics, (2010).
Heller, S., J. Guryan, and J. Ludwig, "Reducing Juvenile Delinquency by Improving Social-Cognitive Skills: Experimental Evidence," (University of Chicago Working Paper, 2012).
Herrera, Carla, Jean Baldwin Grossman, Tina J Kauh, and Jennifer McMaken, "Mentoring in Schools: An Impact Study of Big Brothers Big Sisters School‚ÄêBased Mentoring," Child Development, 82 (2011), 346-361.
Hudley, C., and S. Graham, "An attributional intervention to reduce peer‚Äêdirected aggression among African‚ÄêAmerican boys," Child Development, 64 (1993), 124-138.
Hughes, Erica, and Lindsay Bostwick, "Juvenile justice system and risk factor analysis: 2008 annual report.," (Illinois Juvenile Justice Commission, 2011).
Illinois Juvenile Justice Comission, "Youth Reentry Improvement Report," (2011).
Imbens, G.W., and D.B. Rubin, "Bayesian inference for causal effects in randomized experiments with noncompliance," The Annals of Statistics, (1997), 305-327.
Imbens, Guido W., and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62 (1994), 467-475.
In-Albon, Tina, and Silvia Schneider, "Psychotherapy of childhood anxiety disorders: A meta-analysis," Psychotherapy and Psychosomatics, 76 (2007), 15-24.
Katz, L., J. Kling, and J. Liebman, "Experimental analysis of neighborhood effects," Econometrica, 75 (2007), 83-119.
Katz, L.F., J.R. Kling, and J.B. Liebman, "Moving to opportunity in Boston: Early results of a randomized mobility experiment," Quarterly Journal of Economics, 116 (2001), 607-654.
Kazdin, Alan E., *Conduct disorders in children; Conduct disorders in adolescence; Child Behavior Disorders; Social Behavior Disorders; in infancy & childhood; in adolescence* (Thousand Oaks, CA: Sage Publications, Inc, 1995).
Kendall, P. C. , and L. E. Wilcox, "A cognitive-behavioral treatment for impulsivity: Concrete versus conceptual training with non-self-controlled problem children.," Journal of Consulting and Clinical Psychology, 48 (1980), 80-91.
Kendall, P. C., M. Reber, S. McLeer, J. Epps, and K. R. Ronan, "Cognitive-behavioral treatment of conduct-disordered children," Cognitive therapy and research, 14 (1990), 279-297.
Klein, N.C., J.F. Alexander, and B.V. Parsons, "Impact of family systems intervention on recidivism and sibling delinquency: A model of primary prevention and program evaluation," Journal of Consulting and Clinical Psychology, 45 (1977), 469.
Kling, J. R., J. Ludwig, and L. F. Katz, "Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment," Quarterly Journal of Economics, 120 (2005), 87-130.
Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, "Experimental analysis of neighborhood effects," Econometrica, 75 (2007), 83-119.
Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff, "Economic, neurobiological, and behavioral perspectives on building America's future workforce,"

Proceedings of the National Academy of Sciences of the United States of America, 103 (2006), 8p.
Koegl, C. J., D. P. Farrington, L. K. Augimeri, and D. M. Day, "Evaluation of a targeted cognitive-behavioural programme for children with conduct problems -- The SNAP® Under 12 Outreach Project: Service intensity, age and gender effects on short- and long-term outcomes," Clinical child Psychology and Psychiatry, 13 (2008), 419-434.
Koerner, Kelly, and Marsha M. Linehan, "“Research on dialectical behavior therapy for patients with borderline personality disorder," The Psychiatric Clinics of North America, 23 (2000), 151-167.
Landenberger, N., and M. Lipsey, "The positive effects of cognitive behavioral programs for offenders: A meta analysis of factors associated with effective treatment," Journal of Experimental Criminology, 1 (2005), 451-476.
Larson, KA, and R.W. Rumberger, "ALAS: Achievement for Latinos through academic success," Staying in School. A Technical Report of Three Dropout Prevention Projects for Junior High School Students with Learning and Emotional Disabilities, (1995).
Lee, D.S., "Training, wages, and sample selection: Estimating sharp bounds on treatment effects," Review of Economic Studies, 76 (2009), 1071-1102.
Lee, S., S. Aos, E.K. Drake, A. Pennucci, M. Miller, and L. Anderson, "Return on investment: Evidence-based options to improve statewide outcomes, April 2012 ", (Olympia: Washington State Institute for Public Policy, 2012).
Levin, H., C. Belfield, P. Muennig, and C. Rouse, *The costs and benefits of an excellent education for all of America's children* (Teachers College, Columbia University New York, 2007).
Linehan, Marsha M., Henry Schmidt, Linda A. Dimeff, J. Christopher Craft, Jonathan Kanter, and Katherine A. Comtois, "American Journal of Addition," 8, 4 (1999).
Lipsey, M.W., N.A. Landenberger, and S.J. Wilson, "Effects of cognitive-behavioral programs for criminal offenders," Center for Evaluation Research and Methodology, Vanderbilt Institute for Public Policy Studies, Campbell Collaboration, (2007).
Lipsey, Mark W., "The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic review," Victims and Offenders, 4 (2009), 124-147.
Lipsey, Mark W., and Francis T. Cullen, "The Effectiveness of Correctional Rehabilitation: A Review of Systematic Reviews," Annual Review of Law and Social Science, 3 (2007).
Little, RJA, and DB Rubin, "Statistical Analysis with Missing Data," (2002).
Lochner, Lance, "Education Policy and Crime," in *Controlling Crime: Strategies and Tradeoffs*, Philip J. Cook, Jens Ludwig, and Justin McCrary, eds. (Chicago: University of Chicago Press, 2011).
Lochner, Lance, and Enrico Moretti, "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports," The American Economic Review, 94 (2004), 155-189.
Ludwig, J., "The Costs of Crime: Testimony to the United States Senate Committee on the Judiciary," (Washington D.C., 2006).
Ludwig, Jens, and Douglas L. Miller, "Does Head Start Improve Children’s Life Chances? Evidence from a Regression Discontinuity Approach," Quarterly Journal of Economics, 122 (2007), 159-208.

McCloskey, MS, KL Noblett, JL Deffenbacher, JK Gollan, and EF Coccaro, "Cognitive-Behavioral Therapy for Intermittent Explosive Disorder: A Pilot Randomized Clinical Trial," Journal of Consulting and Clinical Psychology, 76 (2008), 876-886.

McCracken, L. M., and D. C. Turk, "Behavioral and cognitive-behavioral treatment for chronic pain: Outcome, predictors of outcome, and treatment process," Spine, 27 (2002), 2564-2573.

Miller, T.R., M.A. Cohen, B. Wiersema, and National Institute of Justice, *Victim costs and consequences: A new look* (US Dept. of Justice, Office of Justice Programs, National Institute of Justice, 1996).

Moffitt, T.E., L. Arseneault, D. Belsky, N. Dickson, R.J. Hancox, H.L. Harrington, R. Houts, R. Poulton, B.W. Roberts, and S. Ross, "A gradient of childhood self-control predicts health, wealth, and public safety," Proceedings of the National Academy of Sciences, 108 (2011), 2693.

Monahan, K.C., L. Steinberg, E. Cauffman, and E.P. Mulvey, "Trajectories of antisocial behavior and psychosocial maturity from adolescence to young adulthood," Developmental psychology, 45 (2009), 1654.

National Center for Health Statistics, "Health, United States. (With Charbook)," (Hyattsville, MD, 2009).

New York City Independent Budget Office, "The rising cost of the city's juvenile justice system," (2008).

Olds, DL, CR Henderson Jr, HJ Kitzman, JJ Eckenrode, RE Cole, and RC Tatelbaum, "Prenatal and infancy home visitation by nurses: Recent findings," The future of Children, 9 (1999), 44-65.

Orpinas, Pamela, Steve Kelder, Ralph Frankowski, Nancy Murray, Qing Zhang, and Alfred McAlister, "Outcome evaluation of a multi-component violence-prevention program for middle schools: the Students for Peace project," Health Education Research, 15 (2000), 45-58.

Ortmann, R., "The effectiveness of social therapy in prison- a randomized experiment," Crime & Delinquency, 46 (2000), 214-232.

Parsons, Jeffrey T., Sarit A. Golub, Elana Rosof, and Catherine Holder, "Motivational Interviewing and Cognitive-Behavioral Intervention to Improve HIV Medication Adherence Among Hazardous Drinkers," J Acquir Immune Defic Syndr, 46 (2007), 443-450.

Patton, George C, Lyndal Bond, John B Carlin, Lyndal Thomas, Helen Butler, Sara Glover, Richard Catalano, and Glenn Bowes, "Promoting social inclusion in schools: a group-randomized trial of effects on student health risk behavior and well-being," Journal Information, 96 (2006).

Pinker, S., *The better angels of our nature: Why violence has declined* (Penguin Books, 2011).

Puma, M.J., R.B. Olsen, S.H. Bell, and C. Price, "What to Do when Data Are Missing in Group Randomized Controlled Trials. NCEE 2009-0049," National Center for Education Evaluation and Regional Assistance, (2009), 131.

Roderick, Melissa, Jenny Nagaoka, and Elaine Allensworth, "From High School to the Future: A first look at Chicago Public School graduates' college enrollment, college preparation, and graduation from four-year colleges," (Chicago, IL: Consortium on Chicago School Research, 2006).

Rohde, P., G. N. Clarke, D. E. Mace, J. S. Jorgensen, and J. R. Seeley, "An efficacy/effectiveness study of cognitive-behavioral treatment for adolescents with comorbid major depression and conduct disorder," Journal of American Academy of Child Adolescent Psychiatry, 43 (2004), 660-668.

Rumberger, R. W., "Who Drops Out of School and Why." National Research Council, Committee on Educational Excellence and Testing Equity Workshop, "School Completion in Standards-Based Reform: Facts and Strategies (" in *Understanding Dropouts: Statistics, Strategies, and High-Stakes Testing*, A. Beatty, U. Neiser, W. Trent, and J. Heubert, eds. (Washington, DC: National Academy Press, 2001).

Rush, A.J., A.T. Beck, M. Kovacs, and S. Hollon, "Comparative efficacy of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients," Cognitive therapy and research, 1 (1977), 17-37.

Schochet, PZ, J Burghardt, and S McConnell, "Does Job Corps work? Impact findings from the National Job Corps Study," American Economic Review, 98 (2008), 1864-1886.

Schweinhart, LJ, J Montie, Z Xiang, WS Barnett, CR Belfield, and M Nores, "Lifetime effects: The High/Scope Perry preschool study through age 40," (Ypsilanti: High/Scope Press, 2005).

Sexton, T., and C.W. Turner, "The effectiveness of functional family therapy for youth with behavioral problems in a community practice setting," Journal of Family Psychology, 24 (2010), 339.

Shonkoff, Jack P, and Deborah A Phillips, *From neurons to neighborhoods: The science of early childhood development* (National Academies Press, 2000).

Simons-Morton, B, D Haynie, K Saylor, AD Crump, and R Chen, "The effects of the going places program on early adolescent substance use and antisocial behavior," Prevention science: the official journal of the Society for Prevention Research, 6 (2005), 187.

Skye, Dianne Lynn, "Arts-based guidance intervention for enhancement of empathy, locus of control, and prevention of violence," (University of Florida, 2001).

Spelman, W., *Criminal incapacitation* (Plenum Publishing Corporation, 1994).

Swanson, C. B., *Closing the graduation gap: Education and economic conditions in America's largest cities* (Bethesda, MD: Editorial Projects in Education, 2009).

Toplak, M. E., L. Conners, J. Shuster, B. Knezevic, and S. Parks, "Review of cognitive, cognitive-behavioral, and neural-based interventions for Attention-Deficit/Hyperactivity Disorder (ADHD)," Clinical Child and Family Psychological Review, 28 (2008), 801-823.

(www.census.gov/compendia/statab/2010/tables/10s0253.pdf,

Van Voorhis, P., L.M. Spruance, P.N. Ritchey, S.J. Listwan, and R. Seabrook, "The Georgia Cognitive Skills Experiment," Criminal Justice and Behavior, 31 (2004), 282.

Waldron, H. B., and C. W. Turner, "Evidence-based psychosocial treatments for adolescent substance abuse," Journal of Clinical Child and Adolescent Psychology, 37 (2008), 238-261.

Waldron, Holly Barrett, and Yifrah Kaminer, "On the Learning Curve: The Emerging Evidence Supporting Cognitive-Behavioral Therapies for Adolescent Substance Abuse," Addiction, 99 (2004), 93-105.

Walker, J.S., and J.A. Bright, "Cognitive therapy for violence: reaching the parts that anger management doesn't reach," The Journal of Forensic Psychiatry & Psychology, 20 (2009), 174-201.

Weiner, D.A., B. Lutz, and J. Ludwig, "The effects of school desegregation on crime," (National Bureau of Economic Research Cambridge, Mass., USA, 2009).
Western, B., and B. Pettit, "Incarceration & social inequality," Daedalus, 139 (2010), 8-19.
Westfall, P.H., and S.S. Young, *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Wiley-Interscience, 1993).
Wood, A., R. Harrington, and A. Moore, "A controlled trial of a brief cognitive-behavioural intervention in adolescent patients with depressive disorders," Journal of Child Psychology and Psychiatry, 37 (1996), 737-746.
Wooldridge, J.M., "Distribution-free estimation of some nonlinear panel data models," Journal of Econometrics, 90 (1999), 77-97.
WWC Intervention Report, "Twelve Together," (Institute of Education Sciences, 2007).

Table I. Baseline Descriptive Statistics for the Pre-Program Year

	Control Group Mean	Treatment Group Mean	P-value
	N = 1267	N = 1473	
Demographics			
Age	15.70	15.51	0.55
Black	0.72	0.69	0.32
Hispanic	0.28	0.31	0.38
Schooling			
Grade	9.42	9.29	0.25
Old for Grade	0.55	0.51	0.43
GPA	1.68	1.73	0.88
Total Days Present	129.86	133.60	0.35
IEP	0.21	0.20	0.86
Arrests			
Ever Arrested	0.37	0.35	0.36
Number of Arrests for:			
Violent Crime	0.35	0.35	0.56
Property Crime	0.21	0.19	0.96
Drug Crime	0.17	0.18	0.53
Other Crime	0.45	0.47	0.42

Notes: P-values from difference-of-means t-test, adjusted for school fixed effects. The Chicago Public Schools academic year is 170 days, and grade point average is on a 4.0 scale. IEP indicates the presence of an Individualized Education Program as required by the Individuals with Disabilities Education Act. Data sources: Chicago Public Schools administrative data and Illinois State Police arrest records.

Table II. Program Participation

	All Treatment	In-School Only	In- & After-School	After-School Only	Control
Ever Attended	0.49	0.54	0.65	0.21	0.05
Total Sessions Attended	6.64	6.94	9.69	1.94	0.55
Total Sessions | Ever Attended	13.47	12.80	14.97	9.26	11.34
25th Percentile of Attenders	4	5	5	2	3
75th Percentile of Attenders	20	15	22	11	18

Table III. Effect of Treatment on Arrests

Arrest Type	CM	ITT	LATE	Lower-Bound LATE	CCM
			Year 1		
Violent	0.167	-0.0336**	-0.0806**	-0.0521**	0.184
		(0.0165)	(0.0394)	(0.0254)	
Property	0.077	0.0050	0.0120	0.0078	0.066
		(0.0128)	(0.0303)	(0.0196)	
Drug	0.151	0.0026	0.0062	0.0040	0.094
		(0.0178)	(0.0424)	(0.0274)	
Other	0.305	-0.0480*	-0.1151*	-0.0744*	0.320
		(0.0267)	(0.0636)	(0.0411)	
			Year 2		
Violent	0.110	-0.0005	-0.0013	-0.0008	0.09
		(0.0143)	(0.0340)	(0.0220)	
Property	0.057	-0.0032	-0.0076	-0.0049	0.05
		(0.0103)	(0.0245)	(0.0159)	
Drug	0.164	-0.0181	-0.0435	-0.0281	0.17
		(0.0192)	(0.0457)	(0.0295)	
Other	0.264	-0.0417	-0.0999	-0.0646	0.29
		(0.0258)	(0.0614)	(0.0397)	

Notes: Standard errors in parentheses. Baseline covariates included in all regressions. Lower bound uses LATE estimates adjusted for attendance under-reporting. CCM based on main LATE estimate. *** p<0.01, ** p<0.05, * p<0.1

Table IV. Effect of Treatment on School Engagement and Performance

	CM	ITT	LATE	Lower-Bound LATE	CCM
			Year 1		
Index	0	0.0585***	0.1403***	0.0906***	0.218
		(0.0216)	(0.0511)	(0.0330)	
Index Elements					
Days Present	0	0.0450	0.1055	0.0688	0.410
		(0.0280)	(0.0648)	(0.0423)	
GPA	0	0.0595*	0.1312**	0.0885**	0.166
		(0.0305)	(0.0665)	(0.0448)	
Still in School	0	0.0502	0.1205	0.0778	0.147
		(0.0345)	(0.0818)	(0.0529)	
			Year 2		
Index	0	0.0786***	0.1887***	0.1219***	0.039
		(0.0217)	(0.0517)	(0.0334)	
Index Elements					
Days Present	0	0.0470	0.1012	0.0697	0.189
		(0.0350)	(0.0743)	(0.0512)	
GPA	0	0.1012***	0.1999***	0.1456***	-0.225
		(0.0388)	(0.0760)	(0.0552)	
Still in School	0	0.0416	0.0997	0.0644	0.136
		(0.0348)	(0.0826)	(0.0535)	

Notes: All variables standardized on the control group by year, so coefficients are in standard deviation units. Standard errors in parentheses. Baseline covariates included in all regressions. Index elements use only observations with non-missing data on that element. Lower bound uses LATE estimates adjusted for attendance under-reporting. CCM based on main LATE estimate. *** p<0.01, ** p<0.05, * p<0.1

Table V. Sensitivity Analysis of School Engagement and Performance Results

	CM	ITT	LATE	CCM
		Year 1		
Main Results	0	0.0585***	0.1403***	0.218
		(0.0216)	(0.0511)	
Listwise Deletion	0	0.0456*	0.1004*	0.301
(n=2466)		(0.0237)	(0.0516)	
Listwise Deletion, IPW	0	0.0463*	0.1042*	0.306
(n=2466)		(0.0241)	(0.0536)	
Zero Imputation	0	0.0623**	0.1495**	0.258
		(0.0250)	(0.0591)	
CPS Leave Codes	0	0.0509**	0.1222**	0.140
		(0.0214)	(0.0508)	
Multiple Imputation	0	0.0522**	0.1252**	0.254
		(0.0230)	(0.0545)	
		Year 2		
Main Results	0	0.0786***	0.1887***	0.039
		(0.0217)	(0.0517)	
Listwise Deletion	0	0.0667**	0.1316**	0.034
(n=1833)		(0.0305)	(0.0597)	
Listwise Deletion, IPW	0	0.0722**	0.1556**	0.040
(n=1833)		(0.0341)	(0.0728)	
Zero Imputation	0	0.0488*	0.1171*	0.203
		(0.0270)	(0.0639)	
CPS Leave Codes	0	0.0678***	0.1627***	0.080
		(0.0213)	(0.0506)	
Multiple Imputation	0	0.0570**	0.1369**	0.127
		(0.0265)	(0.0629)	

Notes: Regressions use all 2,740 observations unless otherwise noted. Coefficients in standard deviation units. Standard errors in parentheses. Baseline covariates included in all regressions. *** p<0.01, ** p<0.05, * p<0.1

Table VI. Intent-to-Treat Effects by Treatment Arm

	Schooling Index	Violent Crime Arrests	Social Costs of Crime (Murder Trimmed by Half)
		Year 1	
In-School Only	0.0595**	-0.0230	-2,860
	(0.0298)	(0.0228)	(2,636)
After-School Only	0.0683**	-0.0446*	-2,716
	(0.0329)	(0.0252)	(2,918)
Both	0.0505*	-0.0362	-3,275
	(0.0293)	(0.0225)	(2,596)
Control Mean	0	0.167	6884
		Year 2	
In-School Only	0.0709**	0.0080	7,870
	(0.0299)	(0.0198)	(6,224)
After-School Only	0.0906***	-0.0055	-4,785
	(0.0331)	(0.0219)	(6,890)
Both	0.0777***	-0.0053	-3,661
	(0.0294)	(0.0195)	(6,128)
Control Mean	0	0.110	7665

Notes: Social costs of crime use jury award-based estimates from Miller, Cohen & Wiersema (1996), trimming the cost of homicide by half. Standard errors in parentheses. Baseline covariates included in all regressions. *** p<0.01, ** p<0.05, * p<0.1

Table VII. Effect of Treatment on Potential Mediating Mechanisms

	CM	ITT	LATE	CCM
		Year 1		
Switch schools (within CPS)	0.129	-0.0269**	-0.0631**	0.119
(n=2660)		(0.0119)	(0.0276)	
Ever in juvenile justice school	0.043	-0.0076	-0.0182	0.030
		(0.0067)	(0.0161)	
		Year 2		
Switch schools (within CPS)	0.125	-0.0083	-0.0180	0.120
(n=2264)		(0.0133)	(0.0284)	
Ever in juvenile justice school	0.075	-0.0200**	-0.0479**	0.090
		(0.0089)	(0.0211)	

Notes: Coefficients from linear probability models; results from probit analysis are almost identical. Robust standard errors in parentheses. Baseline covariates included in all regressions.
*** p<0.01, ** p<0.05, * p<0.1

Table VIII. Effect of Treatment on Student Survey Measures

	CM	ITT	LATE	CCM
Social-Cognitive Measures				
Grit	0	0.0682	0.1278	-0.072
n = 1,074		(0.0651)	(0.1192)	
Emotional Health	0	0.0703	0.1305	-0.062
n = 1,081		(0.0638)	(0.1158)	
Index	0	0.0700	0.1302	-0.068
n = 1,083		(0.0533)	(0.0970)	
Other Measures				
Academic Press	0	-0.0220	-0.0431	0.005
n = 979		(0.0709)	(0.1348)	
Course Clarity	0	-0.0767	-0.1501	0.067
n = 961		(0.0686)	(0.1305)	
Index	0	-0.0505	-0.0987	0.038
n = 979		(0.0632)	(0.1200)	

Notes: Coefficients in standard deviation units. Baseline covariates included in all regressions.
*** p<0.01, ** p<0.05, * p<0.1

Table IX. Estimated Social Benefits Per Participant

	Low Estimate	High Estimate
	From Realized Crime Reduction	
Savings to Potential Victims	4,613	32,045
	(2,934)	(19,728)
Savings to Government	695*	1,217*
	(406)	(711)
Subtotal	5,309*	33,262*
	(3,017)	(19,804)
	From Potential Increase in High School Graduation	
Earnings Increase to Participant	18,377***	28,441***
	(4,525)	(7,049)
Savings to Government	14,597***	22,591***
	(3,359)	(5,235)
Subtotal	32,974***	51,032***
	(7,833)	(12,205)
	Total	
	38,283***	84,294***
	(8,456)	(23,509)

Notes: All estimates in 2010 dollars. Low-end victim costs from estimates in Miller, Cohen & Wiersma (1996), which include tangible costs (lost productivity, insurance and medical care, etc.) as well as quality of life costs. They trim the cost of homicide by half and use the lower-bound LATE estimate with conservative adjustment for attendance under-reporting. High-end victim costs from Cohen et al.'s (2004) willingness-to-pay estimates and the main LATE estimate. Savings to government are subtracted from Cohen victim costs since they should already be part of the willingness-to-pay estimates. Government savings from reduced crime include arrest, processing, detention, incarceration, diversion, and probation costs. Wage increase and government savings associated with each additional graduate from Levin et al. (2007). Government savings include increases in taxes paid and decreases in public health costs and welfare transfers. Standard errors in parentheses. Baseline covariates included in all regressions.
*** p<0.01, ** p<0.05, * p<0.1

Figure I. Participation Type by Treatment Group

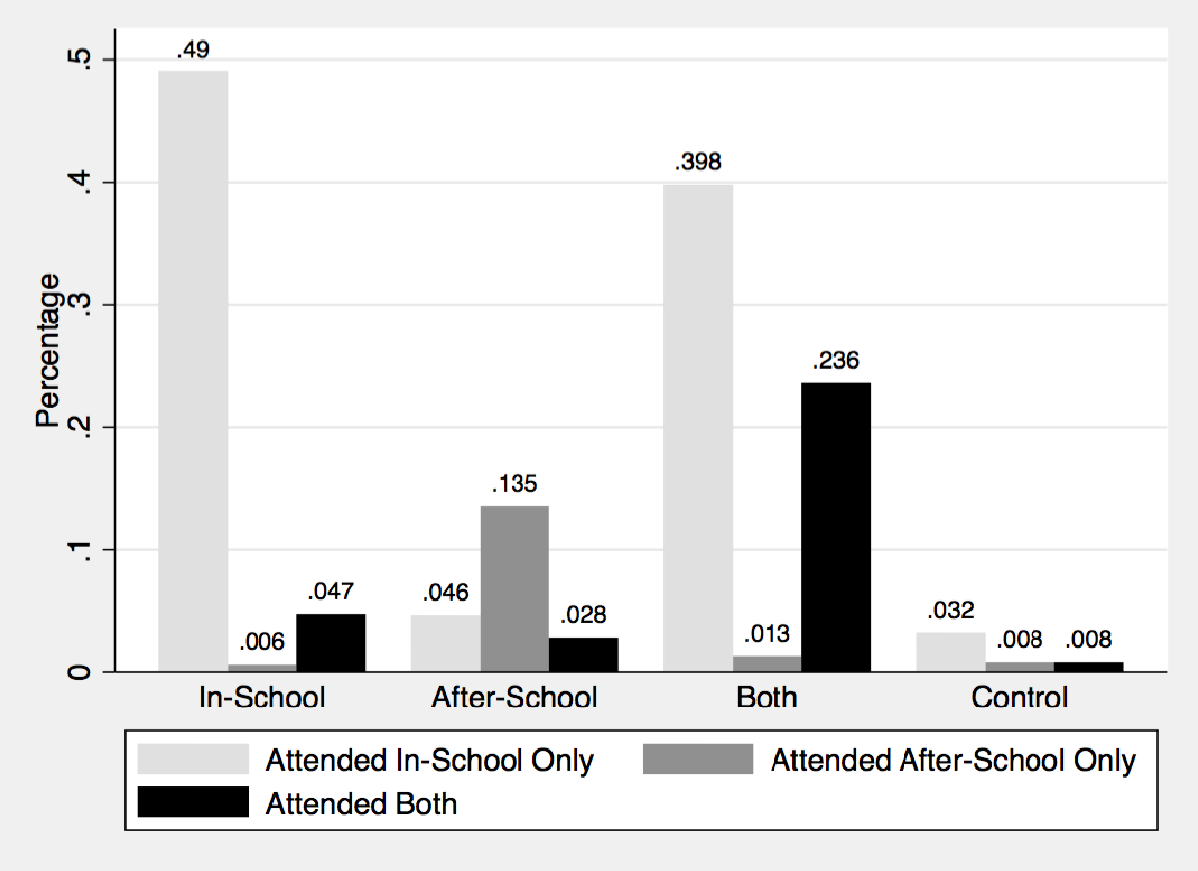
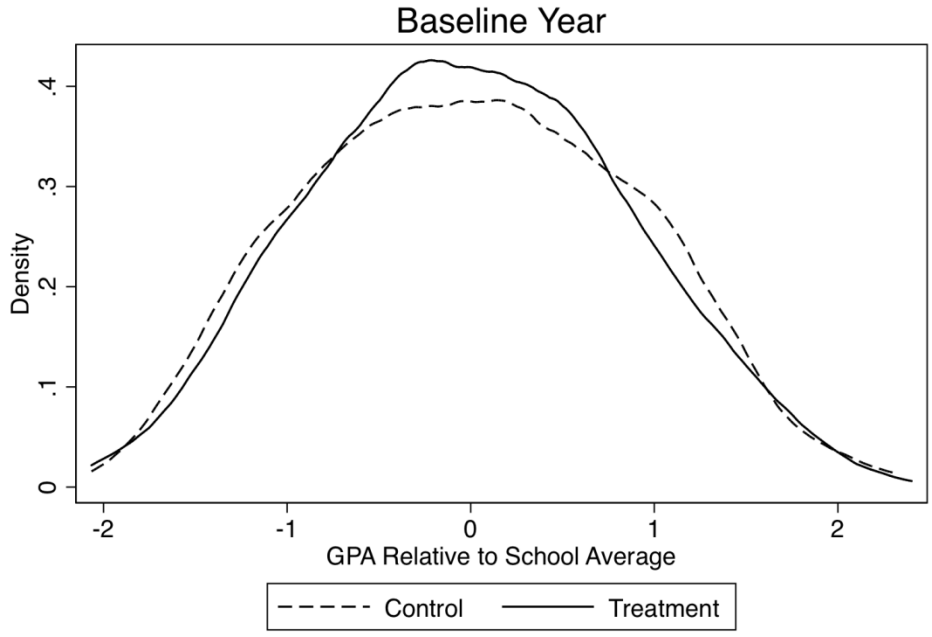
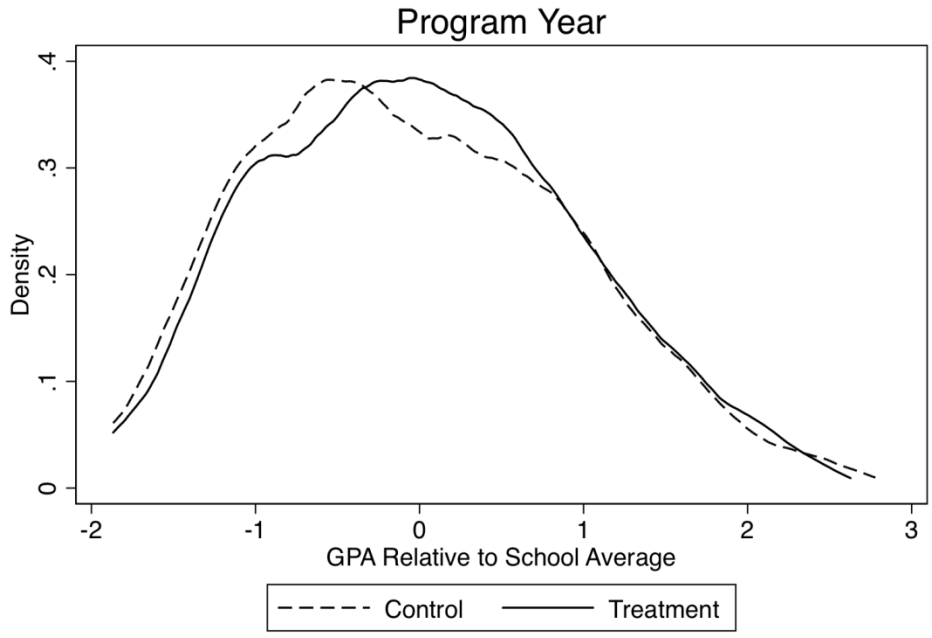


Figure II. Grade Distributions in Baseline and Program Year



Epanechnikov kernel bandwidth .19, Adjusted for school fixed effects



Epanechnikov kernel bandwidth .19, Adjusted for school fixed effects

Figure III. Items Composing Social-Cognitive Skill Measures

GRIT:
To what extent do the following describe you:
I finish whatever I begin
I am a hard worker
I continue steadily toward my goals
I don't give up easily

EMOTIONAL HEALTH:
How much do you agree with the following:
I can always find a way to help people end arguments
I listen carefully to what other people say to me
I'm good at taking turns and sharing things with others
It is easy for me to make suggestions without being bossy
I'm good at working with other students
I'm good at helping people

Appendix Tables

Table A1. Effect of Treatment on School Engagement and Performance, Elements in Original Units

Variable	CM	ITT	CCM	LATE
	Year 1			
Days Present	104.27	2.2144	124.44	5.1966
		(1.3782)		(3.1884)
GPA	1.48	0.0601*	1.65	0.1323**
		(0.0308)		(0.0671)
Still in School	0.88	0.0166	0.92	0.0398
		(0.0114)		(0.0271)
	Year 2			
Days Present	100.16	2.4859	125.56	5.3513
		(1.8477)		(3.9296)
GPA	1.54	0.1026***	1.51	0.2027***
		(0.0393)		(0.0771)
Still in School	0.76	0.0178	0.82	0.0427
		(0.0150)		(0.0355)

Notes. Standard errors in parentheses. Robust standard errors used for linear probability model ("still in school"). Baseline covariates included in all regressions. Table shows variables only for observations with non-missing data on that element. *** p<0.01, ** p<0.05, * p<0.1

.

Table A2. Effect of Treatment on Arrests Using a Lower Match Quality Threshold

Arrest Type	CM	ITT	LATE	CCM
	Year 1			
Violent	0.171	-0.0315*	-0.0755*	0.194
		(0.0169)	(0.0403)	
Property	0.080	0.0063	0.0151	0.070
		(0.0131)	(0.0311)	
Drug	0.154	0.0005	0.0011	0.108
		(0.0181)	(0.0430)	
Other	0.322	-0.0582**	-0.1397**	0.367
		(0.0276)	(0.0657)	
	Year 2			
Violent	0.119	-0.0058	-0.0140	0.114
		(0.0149)	(0.0354)	
Property	0.065	-0.0070	-0.0167	0.070
		(0.0114)	(0.0270)	
Drug	0.174	-0.0226	-0.0543	0.190
		(0.0198)	(0.0471)	
Other	0.272	-0.0371	-0.0889	0.295
		(0.0263)	(0.0626)	

Notes: Standard errors in parentheses. Baseline covariates included in all regressions. *** p<0.01, ** p<0.05, * p<0.1

Table A3. Main Results Assuming In- and After-School Program Effects are Additive

	Schooling Index	Violent Crime Arrests	Social Costs of Crime (Murder Trimmed by Half)
	Year 1		
In-School Only	0.0299	-0.0110	-1,980
	(0.0233)	(0.0178)	(2,061)
After-School Only	0.0333	-0.0303	-1,674
	(0.0245)	(0.0188)	(2,173)
Control Mean	0	0.167	6884

Notes: Results assuming effects of in-school and after-school components are additive. Students assigned to be offered both components have both indicators turned on. Standard errors in parentheses. Baseline covariates included in all regressions. *** p<0.01, ** p<0.05, * p<0.1

Table A4. Individual Treatment Heterogeneity (ITT) by Baseline Characteristics

	Schooling Composite	Violent Crime Arrests
Treatment x No Baseline Violent Crime Arrests		-0.0818**
		(0.0416)
Treatment		0.0302
		(0.0373)
No Baseline Violent Crime Arrests		-0.0423
		(0.0319)
Treatment x Under 1.0 Baseline GPA	0.1130**	
	(0.0512)	
Treatment	0.0261	
	(0.0259)	
Under 1.0 Baseline GPA	-0.3489***	
	(0.0424)	

Standard errors in parentheses. Baseline covariates included in all regressions, excluding variables that are alternative measures of the main effect shown. *** p<0.01, ** p<0.05, * p<0.1

Table A5. Effect of Treatment on Social Costs of Crime in Program Year

	Arrests for All Crimes			
	CM	ITT	LATE	CCM
	Jury Award-Based Costs			
Full Social Cost	8839	-4,995	-11,983	14980
		(3,738)	(8,887)	
Murder Trimmed by Half	6884	-2,976	-7,140	10137
		(1,911)	(4,544)	
	Willingness to Pay-Based Costs			
Full Social Cost	29292	-13,865*	-33,262*	44227
		(8,325)	(19,804)	
Murder Trimmed by Half	25032	-9,465**	-22,707**	33672
		(4,513)	(10,747)	

Notes: All amounts in 2010 dollars. Jury award-based cost calculations use cost of crime estimates from Miller, Cohen & Wiersema (1996). Willingness to pay-based costs use contingent valuation estimates for costs of violent crimes from Cohen et al. (2004). Standard errors in parentheses. Baseline covariates included in all regressions. *** p<0.01, ** p<0.05, * p<0.1