**ONLINE APPENDIX MATERIALS FOR:**

# THINKING, FAST AND SLOW?
# SOME FIELD EXPERIMENTS TO REDUCE CRIME AND DROPOUT IN CHICAGO

September 12, 2016

Sara B. Heller
Anuj K. Shah
Jonathan Guryan
Jens Ludwig
Sendhil Mullainathan
Harold A. Pollack

## APPENDIX A:

## REVIEW OF PREVIOUS INTERVENTION STUDIES

In this section we review what is known from previous studies about the impact of various cognitive behavioral therapy (CBT) programs and related interventions on outcomes for youth, the target population examined in the series of RCTs reported on in our main paper. We then also review studies for both older and younger populations. We divide our literature review up by developmental stage because age-specific factors like maturity, motivation, or developmental plasticity (see, for example, Steinberg 2014) could in principle moderate program impacts. We conclude by noting that remarkably little is known about how to improve schooling outcomes and reduce criminal involvement among disadvantaged teens, even when we look across the entire intervention literature, not limited to a specific intervention strategy.

### I. 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 strategy. Our own inspection of individual studies in this literature suggests that this claim rests largely 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. Focusing on the results of the modest number of previous randomized controlled trials (RCTs) would—if the results are taken at face value—suggest more mixed findings in support of CBT. Yet, as we describe below, many of these experiments also have important methodological limitations, which hinder their ability

to directly explore the impact of CBT on the youth outcomes of greatest policy concern, such as high school dropout or violence involvement.

### *A. Cognitive Behavioral Therapy*

Through the 1970s, most psychological interventions focused on helping people identify and process conflicts and traumas in their pasts. Traditional psychodynamic approaches view presenting symptoms as reflecting fundamental underlying difficulties, which must be addressed before the presenting symptoms can be genuinely relieved (Walker and Bright 2009).

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 psychodynamic approaches, CBT is also more directive, pursuing specific goals such as symptom relief or behavior change, and more structured, focusing on concrete problems and their solutions.

CBT is actually 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 and 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 recognize and regulate their own automatic or impulsive behaviors (Rosenbaum and Ronen 1998). By helping people to think more realistically and effectively, interventions can provide symptomatic relief while ameliorating problematic behaviors. As one manual summarized CBT in criminal justice settings: “By altering routine misinterpretations of life events, offenders can modify antisocial aspects of their personality and consequent behaviors” (Milkman and Wanberg 2007, p.5).

Specific CBT intervention strategies vary, though common elements distinguish CBT from other behavioral interventions (Walker and Bright 2009). 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 the intervention.

In practice, CBT participants are often ambivalent regarding deeply-rooted problematic behaviors, and may be ambivalent regarding continued participation and engagement in the treatment itself. Motivational components and reinforcements are therefore important for successful interventions. CBT is also time-limited. Relatively brief interventions are expected to produce tangible benefits. Most CBT interventions are of 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 (Clarke, et al. 1992; Wood, Harrington and Moore 1996; Brent, et al. 1997; Birmaher, et al. 2000; Rohde, et al. 2004), anxiety disorders (Kendall and Wilcox 1980; Kendall, et al. 1990; Barrett, et al. 2001; In-Albon and Schneider 2007), intermittent explosive disorder (McCloskey, et al. 2008), conduct problems (Kendall and Wilcox 1980; Kendall, et al. 1990; Kazdin 1995; Koegl, et al. 2008), attention deficit hyperactivity

disorder (Toplak, et al. 2008), and emotional dysregulation among severely disordered people (Linehan, et al. 1999; Koerner and Linehan 2000). CBT has also been found to help treat problems like chronic pain (McCracken and Turk 2002), medication adherence (Parsons, et al. 2007), adolescent substance use problems (Waldron and Kaminer 2004; Waldron and Turner 2008), and stress management (Antoni, et al. 2000; Gaab, et al. 2003).

Growing practitioner interest has led to attempts to use CBT to change other problematic youth behaviors. CBT interventions have tried to reduce youth problem behaviors by helping youth to reduce automatic impulses and aggressive behavior, for example through 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 in efforts 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.

*B. CBT Meta-Analyses*

Several meta-analytic reviews conclude that CBT **might** be an effective (and very cost-effective) way to reduce crime and delinquency among both adults and juveniles (Landenberger and Lipsey 2005; Lipsey and Cullen 2007; Greenwood 2008; Drake, Aos and Miller 2009; Lipsey 2009). 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" (p. 23).

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 is typically modest for determining whether average effect sizes systematically differ for RCTs versus non-experimental studies.

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 that it was unrealistic to confine systematic reviews to such studies: "If reviews were restricted to randomized experiments, they would be relevant to only a small fraction of the key questions for policy and practice in criminology. Where there are few randomized experiments, it is expected that reviewers will select both randomized and nonrandomized studies for inclusion in detailed reviews…" (Farrington and Petrosino 2001, p.45; see also the discussion in Greenwood 2008, p.189).

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 experiments that have appeared in the peer-review literature often 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."[1] And even many of the small-scale tests have important limitations, as described next.

## II. RE-EXAMINING EXPERIMENTAL STUDIES OF COGNITIVE BEHAVIORAL INTERVENTIONS FOR YOUTH

We performed our own careful examination of randomized trials 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 and 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, et al. (2011).

[1] 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).

4. Rated a “Top Tier” or “Near Top Tier” intervention by the Coalition for Evidence-Based Policy (www.coalition4evidence.org).
5. Rated a “Level 1” intervention by FindYouthInfo.gov, which was created by the federal government’s Interagency Working Group on Youth Programs
6. Rated “Effective” by CrimeSolutions.gov, sponsored by the U.S. Department of Justice’s Office of Justice Programs
7. 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, which are listed in Appendix Table A.1.

Taken at face value, the previous CBT studies carried out with youth would seem to suggest mixed results for the effectiveness of this intervention strategy. While 12 of the 27 studies find beneficial, statistically significant effects, over half (15 of the 27) find no statistically significant results. However, close inspection of the 27 studies suggest that this pattern of mixed results may be less informative than it first appears, since many of the studies display important limitations that limit internal or external validity as summarized in Appendix Table A.2.

Seventeen of the 27 studies experienced challenges to initial randomization, as suggested either by 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. Of the remaining 10 studies, four studies exhibited attrition rates of at least 20% and/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 and Katz 2005). Of particular concern in the CBT application is the possibility of systematic differences in under-reporting of anti-social behaviors between treatment and control groups. This may be a particular problem for any intervention (like CBT) 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 itself 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 with existing experiments. Six of the 10 successfully randomized studies involved fewer than 100 individuals in the treatment group; three of these six involved treatment groups of less than 50. Small sample sizes are often adequate when exploring common outcomes. Larger samples are required when one explores relatively rare outcomes such as violent offending in community samples of adolescents and young adults. Even when samples are slightly larger, some important studies yield suggestive findings that are not statistically significance due to low power.

Fully 22 of the 27 studies experienced limitations to randomization, attrition difficulties, or relied on self-reported outcomes. Out of the remaining five studies, only one (Dynarski, et al. 1998) included a treatment group exceeding 100 individuals.[2] These challenges were also

[2] The Farrell, Meyer and White (2001) analysis of the Responding in Peaceful and Positive Ways (RIPP) intervention also deserves mention given its similarities to the BAM intervention studied in this paper. Randomized

prominent within the subgroup of 12 studies that found statistically significant benefits to intervention.

Despite the variety of challenges described, several well-executed studies offered methodological and substantive insights that were especially pertinent to the current paper. One of the strongest available studies was performed by Borduin and colleagues (Borduin, et al. 1995). These investigators performed a randomized trial to compare criminal justice outcomes

---

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. Attrited students were older, had lower grade point averages, lower attendance rates, and more out of school suspensions than did students who remained in the study.

Farrell, et al. (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. As noted, 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. Participants displayed a 21.5% attrition rate, with more aggressive students more likely to exit the sample.

Patton, et al. (2006) performed a school-randomized trial in which self-reported anti-social behaviors among 8$^{th}$ 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.

The 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%) are again important limitations.

A Norwegian study by Gundersen and collaborators (2006) provides 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 a 37% attrition rate.

among 176 Missouri juvenile offenders who were assigned to either multi-systemic therapy (MST) (n=92), a more intensive intervention than CBT that also works with family members and others who interact with the youth, or individual therapy (n=84). Participants were reasonably serious offenders. At baseline, they averaged 3.9 prior felony arrests.

Long-term follow-up was subsequently conducted using juvenile and adult arrest records (Schaeffer and Borduin 2005). Over an average of 13 years of follow-up, MST assignment was associated with large and statistically significant declines in post-treatment arrests for both violent and nonviolent crimes. The intervention also reduced periods of incarceration by an average of 62.4 days per year.

Despite these striking results, two aspects of the intervention raise questions regarding generalizability. Only 77 individuals actually received MST treatment. Moreover, treatment was provided by a clinical team of six graduate students in clinical psychology under the direct supervision of the principal investigator. In similar fashion, Timmons-Mitchell and collaborators (2006) obtained significant reductions in rearrests (66.7% versus 86.7%) in an intensive MST intervention that involved one supervisor and 14 therapists serving 48 treatment-group youth. The intensity and small sample size of such studies suggest that these may represent efficacy rather than effectiveness trials in the juvenile justice setting.

The ALAS (Spanish for “wings”) intervention (Larson and Rumberger 1995), bears clear similarities in structure and curricular content to the BAM intervention. Although the study involved a treatment group of only 46 students, treatment and control groups were successfully randomized, with low attrition and the use of 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 and 10 weeks of instruction on problem-solving skills using the ALAS Resilience Builder curriculum. 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 7$^{th}$ grade, were first evaluated at the end of 9$^{th}$ grade, and were again evaluated at the end of 11$^{th}$ grade. These authors found statistically significant improvements in two important measures: student school enrollment at the end of 9$^{th}$ 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 11$^{th}$ 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.

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 12 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, et al. 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 (MRT). This experiment randomly assigned a total of 256 juveniles within a Maryland detention center to a treatment (N=129) or a control group (N=127). The main results in the paper come from analyzing a sample that excluded 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 were not presented. Overall recidivism rates were not different between treatment and control groups; in terms of the number of days to re-arrest, the treatment group had 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 had 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%).

## III. CBT INTERVENTIONS WITH ADULTS

Experiments documenting the impact of CBT with adults—rather than teens—do not add significantly to our understanding of the effects of CBT or related interventions on behavioral outcomes of key policy interest.

Consider, for example, the two CBT experiments with adults that Aos and colleagues (2006) considered to be 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. The parolees were on average 30-years-old and had 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 generally had lower rates of adverse follow-up outcomes—approximately 8–10% of the control mean—for outcomes including 9-month re-arrest rates (38% versus 42%) and 30-month prison re-admission rates (43% versus 47%).

While the point estimates taken at face value suggest some beneficial effect of CBT, the treatment-control differences are not statistically significant; this is at least partly reflective of the limited statistical power of the study. 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 re-arrest/parole revocation ranged from 0.62 to 1.17. Similarly, the 95% confidence interval on the odds ratio for returning to prison ranged from 0.67 to 1.17. 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 and Miller 2009). However, the limited statistical power of the study makes these results tenuous.

Aos and colleagues identified one other randomized experiment for adults—Ortmann (2000)—that they considered to be high-quality. This study reports statistically insignificant treatment-control differences in recidivism equal to about 8–10% of the control mean. Due to the small sample size in Ortmann (2000)—only 111 German prisoners in the treatment group—this study had even less statistical power than Van Voorhis, et al. (2004).

More recently, Blattman, Jamison and Sheridan (2015) presented the results of a randomized experiment with around 1,000 adult men (ages 18–35) in Monrovia, Liberia that involved CBT as a treatment arm. The study used a 2x2 factorial design that also independently randomized some men to receive cash grants. Outcome data came from surveys that asked men to self-report their behavior both at the end of the intervention period and also about a year after the intervention period. The authors used a more intensive qualitative follow-up with a subset of the study participants to try to validate these self-reports. The researchers found that the CBT-only treatment arm reduced anti-social behavior by a statistically significant -0.197 standard deviations in the short-term follow-up surveys and by a statistically insignificant -0.095 standard deviations in the one-year follow-up (standard error 0.080).

## IV. RELATED INTERVENTIONS AMONG CHILDREN

Previous research suggests that early childhood interventions that provide a mix of academic support, parenting training, and other social services between the pre-natal period and age five—such as Perry Preschool, Abecedarian, Head Start, and Nurse/Family Partnership—have long-term effects on educational attainment, employment and earnings, and in some cases, crime (despite documented fade-out of impacts on IQ or achievement test scores) (Currie and Thomas 1995; Olds, et al. 1999; Campbell, et al. 2002; Garces, Thomas and Currie 2002; Schweinhart, et al. 2005; Ludwig and Miller 2007; Deming 2009; Lochner 2011). By eliminating other candidate mediators, researchers have inferred that the effects of these programs on non-academic factors are a key mediating mechanism for the long-term impacts of these interventions. Researchers interpret indirect proxies for "non-cognitive skills," such as teacher-reported behavior and mood, as support for this hypothesis (Heckman, Pinto and Savelyev 2012). However, few studies include good direct measures of these skills.

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 first grade and lasting through high school (Conduct Problems Prevention Research Group 2011). Children in elementary school (grades 1–5) were provided with weekly sessions designed to enhance their social-cognitive skills; the program also provided tutoring to children, parent-training groups, and home visits to work on parenting practices. The intervention became somewhat less intensive during middle and high school. Follow-up studies found that the program did indeed strengthen social-cognitive skills, developmentally appropriate parenting practices, and child behavior during elementary school. But by high school, the intervention no longer had a statistically significant impact on the full study sample; however, there was

evidence of impact for the highest-risk sub-group, but only on outcomes measured by parent-report, not child self-report.

Another study by Hudley and Graham (1993) evaluated an intervention that sought to address hostile intention attribution bias (a social information processing problem in which people have the tendency to assume malevolent intent by others). The researchers randomized a sample of 72 African-American elementary school boys (ages 10–12) who had been screened for problems with aggression. Boys were randomized to the following conditions: 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 a four-month follow-up, the study found some impact of the hostile attribution bias treatment on how boys interpreted intentions of others in constructed scenarios and some impact on teacher reports of aggression, but no detectable impacts on disciplinary referrals at school.

Most recently, a well-executed randomized trial of Stop Now and Plan (SNAP) found evidence of some important behavioral impacts (Burke and Loeber 2015). SNAP randomized high-risk boys ages 6–11 to a program that included several different components for children and families. During the first 12 weeks of the intervention, children were provided a CBT intervention, family counseling, academic supports, and mentoring. Boys assigned to the SNAP condition demonstrated significant reductions in a variety of externalized problem behaviors. At one-year follow-up, youth assigned to SNAP had significantly fewer charges within the juvenile justice system than the control youth.

## V. INTERVENTIONS TO REDUCE DROPOUT, DELINQUENCY, AND YOUTH VIOLENCE

It is possible that our review of the literature may miss key studies, or that we are being overly 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 the criteria for top-tier evidence-based programs by organizations specifically devoted to critically assessing the 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 "positive effects" (two pluses) for school completion (defined as "strong evidence" that the program has a positive effect). 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 about how to reduce youth violence is similarly limited. The influential Blueprints for Violence Prevention reviewed over 1,400 studies, and identified fewer than **ten** "model programs" that were found to reduce crime involvement among teens.

Three of these model programs work with youth already involved with the criminal justice system and are more costly and intensive than the interventions that are the main focus of the present paper (Multi-Systemic Therapy costs $4,500 per participant; Multi-Dimensional Treatment Foster Care costs $27,300 per youth; and Family Functional Therapy costs $1,600–$5,000 per youth).

## APPENDIX B:

## SELECTION OF STUDY SCHOOLS AND STUDENTS

### I. STUDY 1

Figure A.1 shows the elementary and high schools that were included in the two BAM experiments, first in 2009–10 with a sample of elementary and high schools and then again in 2013–15 with a sample of high schools. Three of the high schools were included in both study samples. The figure also makes clear that the schools are disproportionately concentrated in some of the city's most dangerous neighborhoods.

Our team originally recruited 16 CPS schools to participate in study 1. Four of those schools ran separate achievement academies within the same school building. Achievement academies are large, distinct schools for students facing academic or social barriers to conventional academic advancement. Because these achievement academies ran distinct treatment groups and we randomized them separately, we treat them as separate schools—and so we began study 1 with 20 functionally separate schools. The program was never actually implemented in one school, and one school was excluded due to 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.) Since school assignment is a pre-program characteristic, we can separate the sample in this way without compromising random assignment. Our main sample therefore consists of 18 schools.

Five of our study 1 schools are elementary schools, which in Chicago serve students in grades K–8, and the remaining 13 are high schools serving grades 9–12. Study schools were among the lowest-performing in CPS; seven have historically had average GPAs of 2.25 or below (Roderick, Nagaoka and Allensworth 2006). Four were designated "turnaround" schools in the program period, 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 who were 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 Youth Guidance, the non-profit organization implementing BAM, we excluded some students according to their baseline (AY 2008–9) characteristics prior to randomization:[3]

1. Youth who seemed 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 255 students were excluded because they missed more than 60% of days during AY 2008–9 and received a grade of "F" in at least 75% of their courses.
2. Students with specific Individualized Education Program (IEP) designations for serious discrete conditions including autism, speech and language disabilities, "educable mentally handicapped," traumatic brain injury, and diagnosed emotional and behavioral disorders. Service providers determined that youth with these specific diagnoses were

[3] In addition to the exclusions prior to randomization listed below, 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. These students were kept in the study sample to preserve randomization. The choice to not offer the program to youth who were older than grade level resulted in under-enrollment at some locations. A total of 152 age-eligible control youth 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 integrity of the original randomization.

unlikely to benefit from the BAM curriculum. 238 youth were excluded for this reason. Less intensive IEP designations were not used as a study exclusion criterion. Indeed, roughly 20% 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, resulting from some combination of CPS testing schedules (not all grades are subject to standardized testing each year) and student absences during testing days. Because of the large number of missing items for this variable, test scores were not used in sample selection. A total of 929 students were excluded from our initial sample frame because they were frequently absent, satisfied specific IEP designations, and/or generated low risk scores in our selection algorithm.

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 BAM only, after-school BAM-infused sports programming only, both, or neither) within schools (a block-randomized design with schools as blocks).

## II. STUDY 2

After discussions with CPS and the city of Chicago, study 2 focused exclusively on youth in high schools. Youth Guidance received funding from the city and other sources to offer BAM

in nine CPS high schools, some of which were also in our study 1 sample (Figure A.1). These schools are all in the racially and economically segregated south and west sides of Chicago.[4]

In selecting the study sample, our research team followed the study 1 approach outlined above as closely as possible, focusing on male youth in the 9th and 10th grade. We used CPS data on the year prior to random assignment (AY 2012–13) to identify youth, and again excluded youth with IEPs for severe disabilities, as well as those failing over 75% of their classes and missing at least 60% of school days during 2012–13. For study 2 YG also expressed a preference to try if possible to avoid enrolling very old-for-grade students (17 and 18 year olds).

Our team initially obtained a list in early August 2013 from CPS of students the system thought would be attending the nine study schools, and then identified male youth and carried out random assignment to BAM versus control conditions by school and grade (blocks). A sizable share of youth that CPS thought would attend these schools wound up not showing up at their expected school at the start of the academic year, several weeks later. Our team worked with these CPS schools to identify new youth entering the schools who were not on the early August 2013 school rosters, and then randomized these new youth by school and time of school entry. We treat each school-by-grade-by-time-of-entry-period as a separate randomization block in a way that is consistent with how randomization was carried out. In the analysis we keep all students ever randomly assigned in the study sample.

[4] Some readers familiar with Chicago may wonder about the degree to which Hyde Park Academy High School (HPAHS) shares the same level of disadvantage as the other high schools in our sample. The Hyde Park neighborhood is much more affluent on average than its surrounding communities. However, HPAHS is actually located in the Woodlawn neighborhood, a few blocks south of Hyde Park. Woodlawn currently ranks as the 12th most dangerous out of 77 Chicago neighborhoods. HPAHS serves mostly youth in high-poverty areas, and HPAHS's graduation rate is below the CPS average by double-digit percentage points.

## III. STUDY 3

Study 3 arose out of a "natural experiment" that occurred inside the Cook County Juvenile Temporary Detention Center (JTDC), so there was not the same process of selecting study sites as in studies 1 and 2. Our study sample consists of all male youth who entered the facility during our study period. In what follows we provide some additional details about the JTDC facility itself, how this study came about, and how we conducted randomization.

The Cook County, Illinois Juvenile Temporary Detention Center (JTDC), located on the west side of Chicago, is a 500-bed facility that is the largest facility of its kind in the country. In recent years, the JTDC has averaged around 300 youth in residence at any point in time, although historically the average population has been much higher. The JTDC is a pre-adjudication facility for juveniles who are awaiting trial, sentencing, or transfer to a juvenile prison. Compared to the Cook County jail, which holds adult arrestees awaiting trial, the JTDC houses a much smaller share of arrestees who are on average much higher risk.[5] The average length of stay in the JTDC is three weeks, although about 40% of admitted males stay for less than a week, and around 10% of individuals in residence at any point in time are charged as adults in the criminal court and stay for a much longer time (sometimes over a year).

Most of the youth in detention are between 14 and 16 years of age and are disproportionately male, low-income, and either African-American or Hispanic, from high-crime south- or west-side Chicago communities. A research team at Northwestern University led by

[5] In the adult system, arrestees are typically brought before a bond court judge, who sets a bail amount that in turn determines whether the arrestee is detained or released before their case is heard by the courts. In contrast, in the juvenile system in Cook County, upon arrest of a juvenile, police contact the Probation Department's Detention Screening office to administer a risk assessment instrument (RAI) that assigns a score as a function of the severity of the current offense, prior criminal record, and other factors (like whether the youth is on probation or has an open warrant). Most relevant for present purposes is that the juvenile system winds up assigning a **much** smaller (and higher-risk) subset of juvenile arrestees into pre-trial detention compared to adult arrestees in the Cook County Jail (http://jdaihelpdesk.org/specialnotif/Cook%20County%20IL%20Court%20Notification%20Policy%20Manual.pdf).

Linda Teplin has shown that the prevalence of mental health problems among these youth is quite high—particularly for substance abuse disorders (Teplin, et al. 2005).

The JTDC has long been the focus of criticism in Cook County. Reports of ineffective operations, unsanitary conditions, and abuse of youth in detention led to a 1999 federal lawsuit by the ACLU that eventually resulted in a federal takeover of the facility on May 31, 2007 in *Doe v. Cook County*.[6] Earl Dunlap, a nationally-recognized expert in juvenile corrections, was appointed by the courts as the Transitional Administrator in August 2007.[7]

Prior to Dunlap's tenure, the JTDC had been operated as a single 500-bed facility. Youth would spend the mornings attending the Chicago Public School located within the facility, called Nancy B. Jefferson (NBJ). Youth spent most of their time outside of school simply "hanging out" and watching television (Roush 2015). Neither CBT nor behavior modification training were a feature of standard operating procedures in the JTDC prior to Administrator Dunlap's takeover.

One of the innovations implemented under Dunlap's oversight was to divide the facility into 10 essentially separate residential centers of around 50 beds each. From August 2008 to August 2009 the JTDC converted half of these residential centers into what we call "CBT centers" that (as discussed more in the next section) provided youth with group-based CBT sessions, utilized behavior training techniques, required better-educated staff, and incorporated certain therapeutic activities into schooling at NBJ. To enact these reforms, Dunlap reassigned

[6] *American Civil Liberties Union of Illinois* (http://www.aclu-il.org/jimmy-doe-v-cook-county22/ )
[7] *The Cook County Observer* (http://cookcountyobserver.blogspot.com/2009/04/update-on-juvenile-temporary-detention.html)

staff in the facility on some basis other than seniority, which led to a union lawsuit that froze the number of residential centers that were reformed.[8]

Starting November 10, 2009, the JTDC began (after discussions with our research team) randomly assigning male youth entering the facility into one of the four CBT residential units, or to one of the four status quo units (the remaining two units are the admissions units where youth are housed prior to being officially admitted and the medical unit, neither of which offer CBT). Prior to the start of this randomization process, between August 2008 and November 2009, the JTDC intake center staff decided whether youth should go to one of the four CBT residential centers or instead to one of the four status quo residential centers based on some combination of their professional judgment and whether the CBT units had open beds at the time the youth was being assigned to a unit. The random assignment mechanism that started in November 2009 was based on the date of admission to the JTDC together with the youth's date of birth. Specifically, if the youth's day of birth within the month was even and the youth's admission date to JTDC was an even day within the month, OR if the youth's day of birth and admission day to JTDC were both odd, then the youth was assigned to a CBT center. Otherwise, youth were assigned to non-CBT centers. For example, if reviewing intakes on October 15, 2009, the number of the day of intake would be 15, an odd number. For those youth processed on that day, the key variable for them would be the number of the day in their birthday. If a youth's birthday was May 16, 1993, it would be listed as 5/16/93, and the number of the day would be 16, an even number. Since the odd numbered intake date and the even numbered birth date would not match, this assignment would be to a non-CBT center. This assignment algorithm has the advantages of

[8] *American Civil Liberties Union of Illinois* (http://www.aclu-il.org/jimmy-doe-v-cook-county22/) as well as *The New York Times* (http://www.nytimes.com/2011/03/11/us/11cncjuvenile.html?_r=0)

being observable without reliance on admissions' staff reports, and of being nearly impossible to manipulate.

There were some pre-specified reasons for over-riding the treatment placement that was assigned by randomization. Since the primary mission of the JTDC is, appropriately, the safety and well-being of youth, we established five reasons that incoming youth would be classified as an "always-taker" (always receiving treatment) or a "never-taker" (never receiving treatment) after randomization:

1. If the youth was physically, emotionally, or mentally immature he would be housed separately.
2. If the assigned center-type was full when a youth entered the JTDC, the youth would be housed in a center with openings.
3. If the youth had been in the JTDC previously and received CBT treatment, the youth would be assigned to a CBT center. (JTDC staff believed providing youth with continuity in the CBT centers was developmentally beneficial).
4. If there was a safety or related concern that stemmed from the youth's gang affiliation or history of conflict with others, the youth would be placed in a housing unit that the JTDC staff believed was safe.
5. If the youth's scheduled stay in the JTDC was expected to be too short to be assigned to a standard residential center within the facility, the youth would remain within the Alpha, or "admissions," center.

The fact that intake staff could classify youth as always- or never-takers does **not** compromise the strength of our experimental design, as discussed further below, since we can carry out an unbiased intent to treat (ITT) analysis with all the admission spells, regardless of

whether they wound up going into CBT centers or not.[9] We do not exclude those who fall under rules 1–5 from the sample, in part because we do not perfectly observe who is an always- or never-taker under these rules. The inclusion of these youth creates the appearance of considerable non-compliance.

One might worry that the local nature of this effect, which is specific to the population that complied with random assignment in our study, makes it difficult to interpret or generalize results to other settings. In our setting, however, the population of compliers was mainly determined by institutional constraints (e.g., excluding short stays) and administrators' prior beliefs regarding what housing type was appropriate (e.g., safety exclusions for young boys), rather than anything peculiar about the nature of the instrument. If other detention centers were to offer a similar program to only some of their detainees, we might hypothesize that those offered treatment would share many of the characteristics of our complier population. If so, our local average treatment effect (LATE) may in fact be a very policy-relevant parameter.

The random assignment of youth to CBT versus non-CBT centers within the JTDC means that the two types of residential centers within the facility serve populations of youth that should on average be identical.[10] As a result, any differences in post-detention outcomes between the two groups of youth can be confidently attributed to the causal effects of being placed in a CBT center within the JTDC. In February 2011, the litigation between the union and the JTDC

[9] A subset of youth (n = 289) appear in the admissions log but not the housing roster data, which means we are unable to observe if they were housed in a CBT or non-CBT center. These youth are excluded from the sample since we do not observe treatment compliance. Ninety-eight percent of these youth stayed in the JTDC for four or fewer days, and so likely would have been never-takers due to the "short stay" rule.

[10] Note this is true among the population of compliers. Those deemed always- and never-takers due to the five reasons above may sometimes be non-randomly housed with compliers (e.g., those returning to the JTDC more frequently are more likely to have had CBT previously and therefore be assigned to a CBT unit based on reason 3). As such, the peer group in the CBT units is different from the peer group of the controls (older with more previous JTDC spells). Peer composition is thus part of the treatment effect we estimate. We suspect that exposure to older and more frequently delinquent peers would work against any pro-social effects of the treatment itself, such that we would understate the treatment effect that might occur in the absence of this peer sorting.

administration was finally resolved. By March 2, 2011, the remaining non-CBT centers were converted and also started providing youth with CBT, and so randomization ended.

## APPENDIX C:

## ADDITIONAL RESULTS

### I. STUDIES 1 AND 2

Appendix Table A.3 shows that in both of the BAM RCTs around half of youth offered the chance to participate in "treatment" chose to participate. This take-up rate is consistent with other large-scale social experiments (Bloom, et al. 1997; Kling, Liebman and Katz 2007) despite the fact that we randomized first (using administrative data) and then tried to consent youth for program participation, rather than consenting and then randomizing, as is more common.[11] We suspect participation rates for the after-school programming in study 1 are understated 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 was around 13 in study 1 and 17 in year 1 of study 2 and 29 total (study 1 offered more sports sessions but study 2 had sessions over two years).

Appendix Figure A.2 highlights one hard-to-avoid byproduct of running social experiments in challenging circumstances: namely, in study 1 (and to a lesser extent in study 2) there is some treatment-group cross-over (see also Appendix Table A.3). For example, in study 1, 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.

Tables A.4 through A.12 show what happens to our results for studies 1 and 2 when we modify our estimation approach. As noted in the text, the sports participation rates reported by our after-school program provider in study 1 seemed artificially low. This will have no effect on

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

our ITT estimates. But if the take-up rates are under-reported, this will lead us to report overly large estimates for the effects of program participation (since the IV estimate in this case is essentially the ratio of the ITT on outcomes in the numerator to the ITT on participation rates in the denominator, the denominator would be too small).

Table A.4 presents results for BAM study 1 that err on the side of being very conservative: We assume that the sports participation rate in every school is as high as what we see in the school with the highest sports participation rate (70%). We do this by randomly selecting sports non-participants in the two groups offered sports to re-assign as participants. The IV estimates for participation effects in this case tend to be on the order of about two-thirds of the estimates in the main tables, which we consider a lower-bound LATE since this almost certainly overstates the sports participation rate.

In the main paper, we focus on carrying out statistical inference on program effects using data that are pooled from BAM studies 1 and 2, to improve statistical power, and then make various adjustments to the p-values to account for multiple comparisons. We use two methods to test the robustness of our results to testing multiple hypotheses within a "family" of outcomes (for more details see Anderson 2008). The first of our statistical methods 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 re-sampling method to adjust our p-values to account for multiple inference concerns (Westfall and Young 1993). Specifically, we use a bootstrap re-sampling technique that simulates data under the null hypothesis. 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 us to compare the actual set of t-statistics we found 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 re-sampling 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.

A second approach to dealing with inference concerns in the case of multiple comparisons is to control the false discovery rate (FDR), or the proportion of null-hypothesis rejections that are type I errors (Benjamini and Hochberg 1995; Anderson 2008). Within a family of 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 which null hypotheses we can reject at the acceptable q level. We calculate the smallest q-level at which we can reject a program impact on an outcome using two different methods. One is the one-stage procedure from Benjamini and Hochberg (1995). The alternative approach is to use the two-stage FDR-control procedure from Benjamini, Krieger and Yekutieli (2006), which the authors show 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.

Both types of adjustments—to control for either the FWER or FDR—require defining the "family" of outcomes. One issue for family definition is how to handle our measure for total arrests, since it is just a linear combination of our four crime-type-specific arrest measures (violent, property, drug, and "other"). A second issue is whether we consider schooling as a separate family, since it measures distinctly different behaviors than arrests. We handle this in two ways. First, we report the results using five outcomes in our family excluding total arrests—the schooling index plus separate arrest measures for violent, property, drug, and other. We then re-do our analyses with two separate families, using school engagement as a one-element family and then five different crime variables (violent, property, drug, other, and total arrests).

Table A.5 shows that our inferences about the statistical significance of our program impacts on school engagement, total arrests, violent-crime arrests, and arrests for "other" offenses generally do not depend on how we define families of outcomes, whether we control for the FWER or FDR, or whether we use the one-stage or two-stage procedures to control for FDR. The p-values associated with estimates of BAM on school engagement, total arrests, arrests for violent crimes, and arrests for other crimes range from 0.010 to 0.055 with the FWER and FDR controls. It also shows that the two studies' impacts do not significantly differ from each other. (Table A.5 also reports the pair-wise p-values that come from using a permutation test.)

Table A.6 shows that, not surprisingly, the p-values and q-values get somewhat larger if we look at the studies and program years separately, which reduces statistical power by no longer taking advantage of the gains from pooling. What the FWER and FDR values calculated in this disaggregated fashion do not account for is the similarity in results across the two studies. This similarity in findings across studies in itself suggests that the disaggregated impacts reported by study and program year are not just chance findings.

Table A.7 illustrates what happens when we lower the probabilistic-match-quality threshold for what counts as a rap-sheet "match" using the data from BAM study 1. The results are generally similar but slightly attenuated, as we would expect from including more false-positive arrests.

A different potential concern comes from the fact that we are applying OLS to a dependent variable that is the count of youth arrests per year, which is very skewed. To address this potential concern we re-estimate our arrest results using a quasi-maximum likelihood Poisson count data model (Wooldridge 1999). We find that the sizes of the effects in proportional terms are very similar to those obtained from OLS.

Table A.8 shows the program effects on the individual standardized elements of the school engagement index for the program year and the follow-up year for BAM studies 1 and 2. In study 1, during the program year the effects are all about the same size for each element of our index (days present, GPA, and enrollment status at the end of the year[12]) when reported in Z-score terms. During the follow-up year the effect is a bit larger for GPA than the other two elements of the index. We see a qualitatively similar pattern for BAM study 2, where the IV estimates for the effects of participation are roughly of the same magnitude for the elements of the school engagement index in year two, perhaps somewhat larger for GPA than for the other elements. The table also shows the statistical power advantages of using an index that aggregates together individual items with effects that go in the same direction.

[12] Since we only observe the total number of days attended over the entire year, not the timing of those days, we define a youth as enrolled at the end of the school year if he has at least one grade recorded in the 4th academic quarter. The intention is to capture whether youth have remained in school through the end of the academic year. This is not quite the same as "dropout," since youth frequently disengage for periods of time and then return to school. Our measure is not perfect (e.g., we do not observe grades for charter schools in the district). But it avoids relying on staff judgment calls about whether youth are defined as "active" or "inactive" in official enrollment records.

Table A.9 shows the effects by element of the index in their raw units. Using raw rather than standardized units makes it more difficult to compare the size of the different results across outcomes, but has the advantage of making it easier to interpret the magnitudes of each effect.

It is worth noting that while the treatment and control mean values for GPA were similar at baseline, as the first panel of Appendix Figure A.3 shows, in study 1 the variance of the baseline GPA distribution was a bit smaller for the treatment than control group. Specifically, youth assigned to treatment had 0.18 more Cs during the pre-program year ($p = 0.06$) and 0.12 fewer As ($p = 0.13$) than their control counterparts. However, the treatment-control difference in GPA distributions during the program year (shown in the second panel of Appendix Figure A.3) 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 in study 1 the baseline grade distributions were not significantly different at baseline ($p=0.354$), but were statistically different the year of the program ($p=0.032$). In study 2, similar to the test of mean differences in the table, none of the GPA differences were statistically significant (p-values for the Kolmogorov-Smirnov test were: baseline year = 0.540, program year one = 0.776, and program year two = 0.244).

Table A.10 shows what happens to our results when we use different approaches to handle missing data on schooling outcomes. (As a reminder we cannot do any sensitivity analysis on this issue with arrest data because we cannot distinguish people for whom we are missing arrest data from people who were just not arrested). For study 1 our results are qualitatively similar when we use different approaches to 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% during the program year and 33% during the follow-up year, with no missing data on the school enrollment element of the index by construction).

The first row of Table A.10 reproduces our main results for study 1, which follow Kling, Liebman and Katz (2007) 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.[13] MCAR is a testable assumption that seems to fail in our application,[14] perhaps because CPS data on grades and attendance can be missing because youth attend or transfer into low-performing schools with poor record-keeping, transfer to private, charter, or suburban schools, or drop out.

The remainder of Table A.10 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 youths' observable characteristics but not unobserved determinants of outcomes. The results are slightly smaller than our main findings with slightly larger standard errors. The next

[13] While we do control for baseline covariates, that 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.
[14] Regressing an indicator for having non-missing values on all three index elements on baseline covariates produces a global F-statistic of 3.87 ($p<0.0000$) for the program year and 17.32 for the post-program year ($p<0.0000$).

two rows of Appendix Table A.10 use logical imputation.[15] 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 similar estimates (see Appendix D to Heller, et al. 2013 for details).

The bottom panel of Table A.10 shows that the BAM study 2 impacts on school engagement in year 2 (the end of the program period) are, as with BAM study 1, qualitatively similar regardless of what approach we use to handle missing outcome data. Whether the estimate is statistically significant varies somewhat across procedures due usually to slight increases in the size of the standard error, rather than to substantial changes in the size of the point estimate itself.

Table A.11 shows that in BAM study 1, where youth were actually randomized to three separate treatment arms (BAM-only, after-school sports, or BAM plus sports) the ITT effects across treatment arms appear to be fairly similar to one another. The main results reported in the paper pool the three treatment arms together, in part for improved statistical power, in part because treatment-group cross-over makes it hard to learn about specific mechanisms by comparing effects of different treatment arms, and in part because the coaches for the after-school programming received BAM training and were encouraged to deploy those skills during the after-school programming. We focus here on the ITT by arm rather than the IV because instrumenting for "participation" by arm is complicated by the differential under-reporting of attendance across arms.[16] We cannot reject the null hypothesis that the ITT effects are the same across treatment arms. A different approach we tried assumes that the effects of in-school and

[15] 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).

[16] 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.

after-school programming are additive and regresses outcomes against indicators for being assigned to in-school and for being assigned to after-school, so 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%) is too low. Since around half the youth in the other groups participated, for the 21% 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.

In the main paper we exclude motor vehicle violations from our measure of "crime." In Table A.12 we show that the results are qualitatively similar for both BAM studies 1 and 2 if we include them in our arrest measure.

Finally, when we interact different baseline characteristics of youth with treatment assignment in the spirit of exploratory analyses, we generally find few statistically significant differences across identifiable sub-groups of youth in estimated impacts. Figure A.4 presents school-specific ITT estimates for our different program periods (study 1, year 1, as well as years 1 and 2 of study 2), and shows how they relate to the neighborhood homicide rate in the neighborhood served by the school. We examine three outcomes: our school engagement index, total arrests, and arrests for violent crimes specifically.

The top left panel of the figure shows a scatter plot of school specific ITT effects on the school engagement index for year one of study 1. The top middle panel shows the same scatter plot but for year one of study 2, and the top right panel shows the plot for year two of study 2.

The middle panels present school specific ITT effects on total arrests, and the bottom panels present school specific ITT effects on violent arrests.

There is not a strong statistical relationship between neighborhood homicide rates and treatment effects on either total arrests or violent-crime arrests. A univariate regression of the school specific ITT and the school's baseline homicide rate is close to zero for studies 1 and 2, with p-values that never approach usual cutoffs for statistical significance. The correlation between ITT effects on school engagement and baseline neighborhood homicide rates is negative and insignificant (p=0.66) for study 1, and negative and close to traditional levels of significance (p=0.12) for study 2 in year 1.

## II. STUDY 3

In this section we provide a bit more detail about the estimation procedures we use for study 3, which are generally similar to those used for studies 1 and 2 but have a few minor differences, and then we discuss some additional results beyond those presented in the main tables and figures.

As noted in the main text, one slight complication in our JTDC study is that for operational reasons, the randomization algorithm is not binding for all youth in the study sample. Randomization is not binding for those who are physically or emotionally immature, or have been in the JTDC before and were assigned to a CBT unit, or were admitted on a day in which the JTDC's CBT centers were full—in the language of Angrist, Imbens and Rubin (1996), youth in these categories are "always-takers" and "never-takers." Although we do not perfectly observe who the always- and never-takers are in our data, we estimate that the remaining "complier" population should comprise around one-third of the youth who enter the JTDC during our study

period.[17] Our study is therefore akin to an "encouragement design." Note that as long as random assignment increases the probability that some youth get treated, we can estimate both the unbiased causal intent to treat (ITT) effect and the effect of actually participating for compliers, or the local average treatment effect (LATE).

We illustrate our approach in a simple regression framework. Let Z be an indicator for treatment-group assignment through our randomization algorithm, and let D be actual placement in a residential CBT center within the facility. Let some outcome for youth i during or after spell s, like subsequent re-admission ($Y_{is}$), be a function of treatment group assignment, observed variables from administrative records measured at or before baseline ($X_{is}$), and a random error term as in equation (1).

**(1)** $$Y_{is} = Z_{is}\theta_{1is} + X_{is}\delta_1 + \lambda_{1d} + \varepsilon_{1is}$$

The ITT is captured by the estimate of coefficient $\theta_{1is,}$ which has a subscript (i) to make clear the possibility that this effect might vary across youth and (s) to allow for the possibility that treatment has a different effect for repeat spells (although our main estimates constrain this

[17] Our ability to identify which youth are "compliers" is imperfect in that we do not observe all the baseline characteristics that determine always- and never-takers; we only observe the housing units staff place youth into (and staff may not always perfectly comply with the rules). Nonetheless, using this information as an approximation for who the compliers are, we observe that of the 5,728 total male spells during the randomization period, 1,775 never leave the admissions unit (and so would be classified as a short-stay never-taker under rule 5), 1,556 previously received CBT during a prior spell (and so would be an always-taker under reason 3), and 566 are put into the units for small boys, medical concerns, or other "multipurpose" units (a never-taker under rule 1). There are overlaps in these classifications, such that in total 3,425 spells have at least one observable reason to be classified as an always- or never-taker. This would leave 2,303, or 40% of the total sample as compliers. We cannot observe what subset of this population are non-compliers due to safety concerns, nor when units are full (staff aimed to keep the units slightly under capacity, so their definition of "full" is not directly measurable in the housing data). If we suppose another 10% of the remaining youth could fall under these categories, between 30% and 40% of all admissions should have been compliers. We cannot limit the sample to this subset of observed compliers, however, because if staff non-randomly ignored the classification rules, such a sample limitation could undermine the exogeneity of randomization. As such, we include the entire sample in our analysis. The key for our study is that random assignment successfully changed a subset of youths' receipt of treatment in a way that is entirely uncorrelated with observables or unobservables, which is true even if compliance with classification rules was imperfect. As explained below, random assignment to treatment did, in fact, create a large increase in treatment probability among the whole sample.

coefficient to be constant).[18] We condition on baseline characteristics ($X_{is}$) to improve the precision of our estimates by accounting for residual variation in pre-existing conditions, although these covariates are not needed for identification. We include in the estimating equation a series of indicator variables for whether the focal spell is the youth's Nth spell (for different values of N that span the range observed in our data). Since we observe some youth more than once in the data, we cluster our standard errors on individuals.

In our main specifications we also control for day-of-entry fixed effects, $\lambda_d$. With our randomization algorithm, the probability of treatment varies somewhat by day (because there are more odd days in the year, and because with a small number of youth admitted each day, the proportion of youth with even or odd matches will not be exactly constant). Although this variation is small and likely close to random, it is possible that the across-day variation in treatment probability could be correlated with the error term—a potential problem the day-of-entry fixed effects help address. There may also be precision gains to including the fixed effects if, for example, there is seasonal variation in the propensity to recidivate that they help to explain.[19] Our results are generally similar if we drop the day-of-admission fixed effects.

Our main recidivism regressions use a linear probability model (LPM), as the dependent variable is a 0/1 indicator. We also run these regressions using a non-linear probit model as a way to ensure that functional form assumptions are not driving our results. In this setting, it turns out that the average marginal effects from a probit model are very similar to the results from the LPM. Since substantive results do not differ, we use LPM estimates in our main tables (the use

[18] We treat each spell as a separate, though not independent, observation, by clustering our standard errors on the individual. Youth are re-randomized each time they return (although once they have received CBT once, they should become an always-taker due to rule 3).

[19] The same is true of the first-stage in our instrumental variables estimation discussed next: The probability of compliance may vary across days (e.g., depending on how full the CBT or non-CBT units were or on how diligent different admissions staff members were in following the randomization algorithm).

of non-linear models creates additional complications with including hundreds of day-of-admit fixed effects and instrumental variables estimation). We note also that there is essentially no issue of inference in the presence of multiple comparisons in this JTDC analysis (study 3) given our focus on just one or two outcomes.

Since randomization is likely to determine the treatment status for about one-third of our study sample—the potential "compliers" in Angrist, Imbens and Rubin's (1996) terminology—we can also use randomization as an instrumental variable for estimating the effects of actual placement into a CBT center, as in equations (2) and (3).

**(2)** $D_{is} = Z_{is}\theta_{2is} + X_{is}\delta_2 + \lambda_{2d} + \varepsilon_{2is}$

**(3)** $Y_{is} = D_{is}\theta_{3is} + X_{is}\delta_3 + \lambda_{3d} + \varepsilon_{3is}$

Applying two-stage least squares (2SLS) to equations (2) and (3) will estimate a LATE, that is, the average effect on the compliers, as long as (a) treatment only affects outcomes through its effect on CBT receipt (the exclusion restriction) and (b) either $\theta_{2is} \geq 0$ for all (i) or $\theta_{2is} \leq 0$ for all (i) (the monotonicity condition). The fact that only about a third of the sample was "compliers" does not undermine our instrument's relevance: the first-stage F-statistic is 241. The factors determining who is an always- or never-taker do, however, affect the characteristics of those on whom the LATE is identified. We expect the compliers to be those staying in the center for more than a few days, and those who are early enough in their criminal careers not to have been assigned to treatment before (but not so young or physically small to be housed separately).

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 and Liebman 2001). The CCM could differ from the overall control mean if compliers differ from the control

group as a whole (which in our case they almost certainly do, since only an observably different subset of the JTDC population is eligible to be a complier). Katz, Kling, and Liebman's original formulation of the CCM is in a setting where there is no control cross-over, and thus no "always-takers." In order to calculate the CCM in our setting, we leverage the exogeneity of random assignment to assume treatment- and control-group always-takers are equivalent on average (see main paper text and Heller, et al. 2013 for derivation).

In the basic regressions, which use an indicator for readmission to the JTDC as the dependent variable, we discard observations we do not observe for the full follow-up period (18 months for the balanced panel or the number of months in the relevant regression for the full sample), even if we know the youth reoffended within the relevant time period (say, at month 4). This estimates the inverse of a survival probability at a given time (that is, the probability of "failing," or returning to the JTDC, before month *m* rather than the probability of surviving past month *m*) using only uncensored observations.[20]

Table A.13 presents the point estimates and standard errors underlying the figure that was presented in the main text with our results for the JTDC CBT experiment, but now also including the results that **exclude** day-of-admit fixed effects. As noted in the paper's text, these results initially focus on the 2,693 youth for whom we have a full 18 months of follow-up data. The top

[20] Because the data from the JTDC are actually duration data—information on how long each youth remains outside of the JTDC without re-entry—and are right-censored (our JTDC data end on December 21, 2011), one might also want to perform a survival analysis. Survival analysis would allow us to use all the information available by including a censored observation in the analysis up until the point at which it is censored, as well as allow us to estimate other parameters of interest like the treatment effect on the hazard rate. However, hazard models rely on the assumption of ignorable censoring (treatment is uncorrelated with the censoring mechanism), which may be violated in our data because of how Illinois treated juvenile offenders between ages 17 and 18 during the study period. At that time, if a youth committed a misdemeanor, he would be detained at the JTDC (if the offense warrants detention). However, if he committed a felony, he would be automatically sent to the adult system. This fact creates unobserved censoring in our data; we observe the 17-year-olds re-admitted for misdemeanors, but we do not observe whether they commit a felony. Since treatment may change whether or not someone commits a felony, the censoring mechanism here may be non-ignorable. Unless treatment moves youth from misdemeanor to felony offenses, however, we are likely to understate rather than overstate any effect on readmission.

row shows the ITT point estimate and standard error by month from JTDC exit for return rates to the facility, both without and with conditioning on day-of-admit fixed effects. We also present the control mean for these return rates over time. The bottom panel presents the IV estimates for the participation effect, again with and without day-of-admission fixed effects, and now presenting the control complier mean.

The structure of our data are a bit complicated in the sense that youth can have multiple JTDC spells over our study period, so that the results shown in Table A.13 capture an average effect over different numbers of JTDC admit spells. That is, some people will have just one JTDC spell over our study period, while others will have two or three or even four spells. Those people with more spells will contribute more person-spell observations to our main estimates, so their contribution to the overall estimated effect is up-weighted compared to people who have just a single spell. Table A.14 reproduces our estimates using data just from the first spell for each person in our study sample; the point estimates are qualitatively similar to those reported in Table A.13, though less precise in part due to the smaller sample size.

The results in Tables A.13 and A.14 looking at whether youth are re-admitted to the JTDC hint at the possibility that the absolute magnitude of the effect peaks around 14 months and then declines a bit thereafter, although this could be an artifact of the fact that a growing share of youth are re-admitted over time. That is, the share of youth who are still at-large in public (and so "at risk" for being readmitted) declines over time, which could have a mechanical effect in shrinking the size of the point estimate.

To address this possibility, in Table A.15 we now estimate the effect of the random assignment and of actual CBT participation on the **number** of re-admissions to the JTDC facility by time since exit from the facility. This comes from re-estimating equation (1) above with a

count variable now as the dependent variable. We count all readmissions within the follow-up period, even if youth are randomly assigned to a different treatment status in later spells. These effects therefore capture everything that follows release, including later treatment. Table A.15 shows that the effects on the number of JTDC readmissions no longer declines in magnitude towards the end of our study period, and viewed over the full follow-up period tends to grow in magnitude.

Table A.16 replicates the analysis using the combined count of the number of JTDC re-admissions and the number of arrests that youth experience as the dependent variable. The size of the IV estimates are not very different from those reported in Table A.15, which looked just at return rates, but the standard errors are now larger.

Tables A.17 through A.20 replicate the results presented in Tables A.13–16 but now using data for all n=5,728 male admissions to the JTDC over the course of our study period (that is, all youth who were randomly assigned to CBT units or control units). Each column presents the results of a different follow-up period, and in each case we use only the observations that we observe for the entire period in that column.[21] With this "imbalanced panel" the effects we estimate during the early period following random assignment are now more muted than what we saw with the balanced sample (that is, the sample of youth for whom we have data for all 18 follow-up months). The sources of treatment heterogeneity underlying this pattern are a topic we will focus on in additional research with these JTDC data in the future. Expanding the sample size like this winds up providing some additional statistical power, so that for example when we

[21] There are two observations in the full sample used in Table A.20 that are missing arrest data. This occurred because after the matching process, we discovered that these two observations had been matched to individuals in the Illinois State Police records who had already been matched to different study youth (i.e., two sets of arrests records were matched to four different study youth). Manual review of all the data on the youth suggested that the double matches were not the same person, but rather that the second matches were false positives. Although by definition this means that there were no higher-quality matches for these youth in the police records, we did not continue searching for matches once the initial match was made. As such, we treat these observations as missing rather than assuming they had no arrests.

look at our noisiest outcome (the number of combined returns to the JTDC and arrests) we now see some point estimates that are statistically significant (Table A.20).

## III. POOLED RESULTS

Table A.21 reports results from pooling all three studies' data together. The advantage of pooling data from all three studies together is to further improve the statistical power available for detecting impacts of programs that we hypothesize operate through similar underlying channels, and to test whether the effects of each intervention are similar to one another. However it should be kept in mind that the outcome variable is measured in different ways across the three studies, partly because the units of measurement are different (arrests in the BAM studies 1 and 2, and a combination of arrests and return rates to the JTDC in study 3) and partly because study 1 captures a mix of program and post-program data, study 2 only captures arrests during the program, and study 3 only captures behavior post-program. In addition the compliers in study 3 may be quite different from the compliers in the BAM studies. These differences may complicate somewhat the interpretation of the pooled coefficients.

The dependent variable in Table A.21 is the number of observed criminal incidents per youth per year. For study 1, that is simply the total number of arrests during year 1 (measured in state police arrest records). For study 2, it is the total number of arrests during the program period scaled to an annual rate (19 months of data / 19 * 12, measured in city police department arrest records). For study 3, we use the total number of re-admissions and re-arrests (at the state level) observed in the 12 months after release. This variable will be higher for study 3 both because it counts an additional outcome (re-admissions) and because the youth in the JTDC are more criminally active than the youth in the BAM studies. Because some youth are in multiple studies, we cluster our standard errors on individual. Not all the baseline covariates used in the

BAM studies are available in the JTDC data, and vice versa; we impute zeros for missing covariates and include a dummy variable to indicate the covariate is missing.

Panel A shows that across studies, participants average about 0.24 fewer incidents in a year, a 24 percent reduction relative to the control complier mean. Panel B breaks this out by study. The first row is the main effect for study 2, followed by interactions of treatment with dummies for studies 1 and 3.[22] Although the coefficient on the interaction term for study 3 suggests that the LATE is about twice as large in absolute magnitude as in study 2, the difference is not statistically significant. In proportional terms, the study 3 treatment effect is actually smaller than in the BAM studies (12 percent for study 3 versus 30 and 26 percent for studies 1 and 2). We fail to reject the null that the program effects are the same across the studies (the p-value at the bottom of the panel is from a joint test of whether both interaction terms are 0).

Panel C adjusts for the fact that the outcomes and youth populations are different across studies by standardizing the dependent variable within each study on the control group (that is, subtracting off each study's control mean and dividing by that study control group's standard deviation). The point estimates are all within 0.04 standard deviations of each other and not significantly different, although the standard errors are somewhat large.

## IV. MECHANISM RESULTS

In the main text we present estimates that show how much of the overall intervention effect (P) on outcome Y could be explained by the intervention's effects on the candidate mechanisms (M) captured on the CCSR surveys of youth in BAM study 1. We run a non-experimental regression using just data from the randomized control group to estimate the relationship between the mechanism and the outcome (M→Y), multiply that by the

[22] The main effects for each study are absorbed by the randomization block fixed effects. The control means for each study are in the row that reports the effect (or interaction) for that study. If we also allow the effects of all the other baseline covariates to vary by study, we gain a tiny amount of precision, but the results are basically identical.

experimentally-estimated effect of BAM on that mechanism (P→M), and then divide that by the experimentally-estimated effect of BAM on the outcome (P→Y). The obvious limitation of this approach is that our regression estimate for the M→Y relationship could be confounded by other omitted variables that lead to a bias that is difficult to sign. The main text presents results for our schooling outcome during the program year and our violent-crime arrests outcome. Tables A.22 and A.23 show results for several additional outcomes as well, and (consistent with the main results presented in the text) suggest that the candidate mechanisms captured by the CCSR survey can usually account for only a modest share of the BAM effect on the different outcomes.

Table A.24 presents an overview of the design of our iterated dictator game experiment that we carried out with a sample of youth from BAM study 2. We expected that participants who had previously been assigned to BAM would make slower, more deliberate decisions than participants who had been assigned to a control condition—that is, that BAM would cause youth to notice that they are feeling angry after having been provoked, but then they would pause and try to make sure they are construing the situation correctly, considering whether this a "code of the streets" situation where provocation might require retaliation, or a "code of the school" situation where being challenged requires compliance. We most clearly expected this prediction to hold for the first "no delay" condition of our experiment, where youth were asked to go ahead and report how much they would take from the other student without any delay or distraction induced by our research team. Conditions 2–4 were intended to attenuate the BAM-control difference by providing a small temporary dose of "slowing down" to the controls (conditions 2 and 3) or by trying to undo the reflection process of those in the BAM group (condition 4).

Table A.25 presents additional results for our decision-making experiment that we carried out with youth in study 2 to measure the effects of BAM on automaticity. The main table reports

results for the log of the time youth took to make a decision about how much to take from the other "study subject" in our iterated dictator game. We use the log value as our main result (also reproduced in the right-hand column of Table A.25) because of the skew in this measure. Table A.25 shows that the results are qualitatively similar when we use decision time as measured in raw units (seconds), or raw units and exclude outlier values.

Figure A.5 shows the cumulative distribution functions (CDFs) for decision-making time for youth randomly assigned to the BAM versus control groups. Because the distribution is so skewed, we present the results three different ways: first showing the CDFs for actual decision-making time for the full distributions; then in the middle panel, in order to make it easier to see the part of the distributions where there is most of the separation, we show the CDFs truncating both distributions at 20 seconds; and in the bottom panel we show the distributions for the log of decision-making time.

Table A.25 also presents results that combine data from our different experimental conditions in different ways. Our theory of automaticity has the cleanest prediction about what we would expect to see under condition 1, where youth in our study sample were asked how much they would like to take from their "partner" in our iterated dictator game experiment without any special testing conditions inserted or applied. Our theory predicts that BAM should get youth to slow down and reflect on what they are doing, which should in turn be reflected by an increase in the time it takes to make a decision about how much to take (and whether to retaliate)—which is what we see in the data. Our theory of automaticity has a less clear prediction about what we would expect to see in the other conditions, since in condition 2 the decision-making exercise itself tried to explicitly get all youth (BAM and control) to do at least part of what we expected BAM to get youth to do on their own (slow down) while in condition 3

the decision-making exercise tried to get all youth to do both things we expected BAM to get youth to do on their own (both slow down and reflect). Condition 4 got all youth to slow down and then tried to intentionally undo any reflection that youth in BAM might do, by trying to get them to ruminate on their partner's behavior—so this condition in some sense tried to "un-CBT" the BAM youth for at least part of how we think CBT works. The main text focuses on the condition 1 results but also shows the results from pooling data from all four conditions. Table A.25 shows that compared to the results from pooling all four conditions together, the results are similar when we focus just on data from conditions 1–3 (excluding data from condition 4).

One potential concern is the possibility that these results are somehow an artifact of response rates that are substantially less than 100%. To examine this possibility we take advantage of the fact that random assignment was carried out within schools, so each school is essentially its own experiment. In Figure A.6 we plot the school-specific BAM impact on the log of how long it took each youth to decide in our decision-making experiment against the response rate for our decision-making exercise in that school. That is, if (s) indexes schools, we estimate equations (2) and (3) from the main text separately for each school to get $\pi_{2s}$ and also calculate school-specific response rates $R_s$ to our decision-making task. Figure A.6 plots $\pi_{2s}$ against $R_s$ for schools in our study sample; the line in Figure A.6 shows the slope from running the regression:

(4) $\pi_{2s} = \rho_0 + \rho_1 R_s + v_s.$

The figure is calculated by pooling data from all 4 conditions in our decision-making experiment to maximize sample size. The figure suggests the BAM effect on this candidate mechanism was, if anything, **larger** in schools with higher response rates. That is, to the extent to which low response rates leave more room for imbalance in the compositions of treatment versus control group respondents, we do not see any attenuation in the treatment-control

difference in our outcome (decision-making time) as we look in schools with less room for this type of imbalance (i.e. higher response rates).

## V. BENEFIT-COST ANALYSIS

Any sort of benefit-cost analysis for a social program like this is necessarily speculative and subject to a number of caveats. Accurately measuring all program costs can be difficult, particularly if some of the program inputs are donated "in kind." Measuring program benefits can be complicated by the fact that the outcomes we measure, such as arrests, do not always correspond one-to-one with the underlying behavior that affects societal well-being, such as criminal behavior. In addition the benefits of some outcomes such as increased high school graduation only accrue long after our study period ends, so we cannot directly observe them. Monetizing many of these benefits is challenging. Nonetheless we present some back-of-the-envelope calculations as a way to roughly benchmark the scale of the social benefits in relation to the costs of this intervention strategy.

Table A.26 presents the results of our basic benefit-cost analysis for BAM study 1. The table breaks the benefits of the program in two parts: the benefits from the realized crime reduction during the program year (both direct savings to the criminal justice system and the broader benefits to society from reduced crime) and—more speculatively—the future benefits from increased graduation. Since costs are more natural to think of in per participant terms (rather than per randomly-assigned youth), the table shows IV estimates, with monetary estimates of program benefits as dependent variables, for youth who chose to participate in the program.

We focus on BAM study 1 for this exercise because it is the one for which we have the longest-term data on benefits. For BAM study 2, we only have data that covers outcomes in the

program period, so we cannot measure follow-up impacts—including subsequent high school graduation. For the JTDC study, our outcomes are specific to spells, not youth. This generates conceptual complications in estimating effects on long-term outcomes like high school graduation when most youth eventually become treated (receive CBT during a JTDC spell).

There are some unavoidable sources of uncertainty and measurement error in the data we have available to estimate the dollar-value benefits of BAM in study 1, and so our approach is to try to provide reasonable upper- and lower-end estimates (the left and right columns of Table A.26). One difference between the columns has to do with how we define the participation rate for BAM. Since the IV is essentially the ITT divided by the participation rate (technically the difference in participation rates between youth assigned to treatment versus control groups), the smaller the participation rate we use, the larger the IV estimate. As explained in the main text, we suspect that participation rates for the sports programming are understated. For the upper-bound benefit estimates in Table A.26, we therefore use the sports program participation rate data as reported (potentially too small) to calculate the IV. The lower-bound benefit estimates in the table instead uses our upper-bound estimate for the sports participation rate (likely too big, see discussion above and Table A.4).

To calculate the direct cost of each arrest to the criminal justice system, we use the cost of processing at the police station and the costs of later stages of punishment for the subset of arrested youth who experience them (detention, incarceration, probation, etc.). We know of no national estimates for these costs, so we construct them 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 (New York City Independent Budget Office 2008; Hughes and Bostwick 2011; Illinois Juvenile

Justice Commission 2011). We find that the cost of an average arrest to the criminal justice system is between $5,770 and $6,524. The range comes from varying estimates of the cost of juvenile detention in Chicago; we use the lower (higher) number for the lower- (upper-) end estimates and vice versa. This is likely a conservative estimate given that we do not incorporate court and policing costs; a similar calculation for North Carolina found each arrest costs an average of $7,300 (Governor's Crime Commission 2009). We multiply these costs times the number of arrests for each youth in year 1 and use the total as the dependent variable in the "savings to government" row of the crime reduction benefits.

Monetizing the broader social costs of crime—the outcome in the "savings to potential victims" row—is not a straightforward exercise. Conceptually, the ideal way to measure the value to society from reductions in future crime 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 approach 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 Cook and Ludwig 2000; Cohen 2005). An additional practical challenge is that the statistical value of life adds a huge amount of variance by assigning a very high cost to a very small number of fatal crimes.

Our approach is to make one set of choices that errs on the low side of these issues and another set of choices that errs on the high side, as a way to demonstrate how much these choices

matter. Specifically, for the lower-end benefit estimates in the table, we start by following the basic strategy in Kling, Ludwig and Katz (2005), assigning each type of crime the social costs[23] estimated by Miller, Cohen and Wiersema (1996), inflated to 2010 dollars, that rely on jury award data. We use the same estimates for crimes not included in the Miller, et al. estimates as Kling, Ludwig and Katz (2005). To be conservative in terms of the statistical value of life, we (arbitrarily) divide the costs of homicide by half. The upper-end figures instead use the willingness-to-pay estimates for the costs of each type of violent crime from Cohen, et al. (2004), which are substantially higher than those from Miller, et al. (1996), and use the estimated social cost of homicide as given with no adjustments. Since Cohen, et al. (2004) only estimate willingness-to-pay for a subset of violent crimes, we use the Miller, et al. estimates for the remaining crime types. Note that to the extent survey respondents incorporate the benefits to the government in their responses about willingness-to-pay for crime reduction (e.g., from their own reduced tax burdens), this approach could double-count some of those benefits. We do not incorporate any other benefits to offenders or their families of reduced criminal justice involvement, although there is reason to believe that less incarceration could have substantial benefits for them (Aizer and Doyle 2015; Mueller-Smith 2015).

For each individual in our study, we multiply the victim cost of each arrest by the number of arrests for that crime during year 1, then sum all the costs to obtain an overall victim cost-of-crime for each person in the study. That cost is the dependent variable in the IV regressions reported in the "savings to potential victims" row of Table A.26. Since these crimes all occur during the program year, we do not discount them. Our estimates are based on arrests, and not all criminal offenses result in arrest. So we may understate the total social benefits from averted

[23] Note that the table reports social benefits, but the behavior we are measuring—crime—generates costs. In practice, we assign negative numbers to the dollar values associated with each crime, so that the positive coefficients in the table represent the (positive) benefits of reduced crime.

crimes among treatment youth (although it is also true that not all arrested youth actually committed the crime for which they were arrested).

As the left column of Table A.26 shows, using the upper-bound participation rate and lower end costs of crime, we estimate that benefits from reduced criminal behavior during the program year alone outweigh the program costs by about 5-to-1. If we instead use participation as reported and the higher valuations of crime (right column), benefits may outweigh program costs by as much as 30-to-1.

The bottom panel of Table A.26 estimates the potential future benefits from increased high school graduation. Unlike the concurrent crime benefits in the top of the table, the future benefits from graduation have not yet accrued for our study youth. As such, we consider our estimates of the benefits from graduation considerably more uncertain than the crime benefits, which are realized during our study period and so can be directly measured.

To calculate these benefits, we focus on two key outcomes where the literature provides at least some arguably causal evidence on the magnitude and social benefits of graduation effects: earnings and health benefits to the participant. Our calculations involve many simplifying assumptions and are intended to give a sense of the possible magnitude of a few key benefits, not an exhaustive benefit-cost analysis. Our first simplifying assumption is that each graduate accrues one additional year of education relative to each non-graduate. This makes it easier to apply estimates from the education literature, since many studies focus on the returns to an additional year of schooling rather than the sheepskin effect at graduation. Using the Chicago Public School data to measure how many years of schooling are actually completed by BAM study members is complicated, in part because we do not know exactly when during a school

year someone stopped attending. Given that most dropouts leave school prior to 11th grade, we suspect this is a conservative decision that all else equal leads to understating benefits.

The causal effect of education on future earnings is a matter of much debate (see, e.g., Card 1999; Heckman, Lochner and Todd 2006).[24] Across different empirical strategies and populations, estimates of increased earnings due to an additional year of high school (or graduation) tend to fall in the range of 8-12 percent (Card 1999). As such, we use an 8 percent increase in lifetime earnings as our low-end estimate and 12 percent as our high-end estimate. To the extent that youth on the margin of dropping out—as the additional graduates in our study are—experience larger returns, and to the extent that improvements in observed wages may not fully capture increases in the probability of being employed, this may understate the earnings benefits of graduation. On the other hand, to the extent that returns to high school have been falling over time (Goldin and Katz 2009), our extrapolation of past estimates into the future may be overly optimistic.

As a baseline from which to calculate the earnings increase, we use estimates of median lifetime earnings for male high school dropouts by race from the Census Bureau (Julian and Kominski 2011).[25] We sum the discounted additional earnings over 40 years. To monetize the health benefits of education, we use estimates from Cutler and Lleras-Muney (2008), who suggest the present value of decreased mortality from an extra year of education is between

[24] Estimates vary by how authors define "returns" to schooling; how they deal with selection bias, measurement error, and the inability to observe wages for non-workers; and which population is driving identification given heterogeneous treatment effects.

[25] These estimates come from the Census Bureau's synthetic cohort analyses, and total $810,681, $776,007, and $128,997 for Hispanic, white, and black male dropouts respectively. Since these estimates are in 2008 dollars and the inflation index is only $1.01 between 2008 and 2010, we do not adjust for inflation. To simplify the discounting of these 40-year totals, we ignore the curvature of age-earnings profiles and just assign each year 1/40th of the total. We then discount at 3 percent for the upper-bound estimates and 5 percent for the lower-bound estimates to calculate the present value of the earnings benefit from graduation, assuming youth start earning 5 years in the future (the Census earnings estimates start at age 25).

$13,500 and $44,000.[26] We use the sum of the earnings and health benefits as an estimate of the benefits of graduation to the participant.

To calculate the coefficients reported in the table, we multiply these graduation benefits by an indicator for whether each youth graduated, and use the result as the dependent variable in an IV regression. The lower-end regression uses the measure of graduation that generates the smallest treatment effect (treating transfers as dropouts); the upper-end regression uses the measure with the largest effect (treating transfers as graduates). The "cost of additional schooling" row assigns the undiscounted instructional cost of the extra year of schooling to each graduate ($7,946 for Chicago Public Schools in 2010) (Illinois State Board of Education 2015).[27] We note that there are likely a number of benefits of increased graduation to the government that we do not count against this cost, including increased tax revenue from higher earnings, decreased crime in adulthood, and reduced public service use.

At the lower end, our conservative estimates for the future benefits of graduation are relatively low (just over $700 and not statistically significant). But at the high end, they add about $5,000 of additional benefits. Combining both the realized crime benefits with the speculative graduation benefits results in benefit-cost ratios for the program overall that are between 6-to-1 and 36-to-1.

[26] Although these estimates come from OLS regressions that are not explicitly causal, their paper finds that more plausibly causal estimates from various IV approaches are not that different from OLS estimates. Additionally, Lleras-Muney (2005) suggests that the causal effects for our study population—males at the lower end of the distribution—may be even bigger than OLS estimates.

[27] This is the average per-pupil cost of teaching that the state reports for the Chicago Public School district. It does not include capital expenditures, summer school, or other operational expenses. We make this choice because the cost of additional teaching seems closer to the marginal cost incurred for an extra year of schooling for one student. If the program were implemented at a large enough scale to require additional school buildings or infrastructure to handle the additional students, this cost may be higher.

REFERENCES

"Doe V. Cook County," in *F.3d*, (United States Court of Appeals, Seventh Circuit, 2015).

Aizer, Anna, and Joseph J Doyle, "Juvenile Incarceration and Adult Outcomes: Evidence from Randomly Assigned Judgest," Quarterly Journal of Economics, 130 (2015), 759-803.

Alexander, James F, and Bruce V Parsons, "Short-Term Behavioral Intervention with Delinquent Families: Impact on Family Process and Recidivism," Journal of Abnormal Psychology, 81 (1973), 219.

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

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin, "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91 (1996), 444-455.

Antoni, Michael H, Stacy Cruess, Dean G Cruess, Mahendra Kumar, Susan Lutgendorf, Gail Ironson, Elizabeth Dettmer, Jessie Williams, Nancy Klimas, and Mary Ann Fletcher, "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), 29-37.

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

Armstrong, Todd A, "The Effect of Moral Reconation Therapy on the Recidivism of Youthful Offenders a Randomized Experiment," Criminal Justice and Behavior, 30 (2003), 668-687.

Barrett, Paula M, Amanda L Duffy, Mark R Dadds, and Ronald 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.

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

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli, "Adaptive Linear Step-up Procedures That Control the False Discovery Rate," Biometrika, 93 (2006), 491-507.

Birmaher, Boris, David A Brent, David Kolko, Marianne Baugher, Jeffrey Bridge, Diane Holder, Satish Iyengar, and Rosa Elena Ulloa, "Clinical Outcome after Short-Term Psychotherapy for Adolescents with Major Depressive Disorder," Archives of General Psychiatry, 57 (2000), 29-36.

Blattman, Christopher, Julian C Jamison, and Margaret Sheridan, "Reducing Crime and Violence: Experimental Evidence on Adult Noncognitive Investments in Liberia," (Cambridge, MA: National Bureau of Economic Research, Working Paper No. 21204, 2015).

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

Borduin, Charles M, Barton J Mann, Lynn T Cone, Scott W Henggeler, Bethany R Fucci, David M Blaske, and Robert A Williams, "Multisystemic Treatment of Serious Juvenile Offenders: Long-Term Prevention of Criminality and Violence," Journal of Consulting and Clinical Psychology, 63 (1995), 569.

Brent, David A, Diane Holder, David Kolko, Boris Birmaher, Marianne Baugher, Claudia Roth, Satish Iyengar, and Barbara Johnson, "A Clinical Psychotherapy Trial for Adolescent Depression Comparing Cognitive, Family, and Supportive Treatments," Archives of General Psychiatry, 54 (1997), 877-885.

Burke, Jeffrey D, and Rolf Loeber, "Mechanisms of Behavioral and Affective Treatment Outcomes in a Cognitive Behavioral Intervention for Boys," Journal of Abnormal Child Psychology, (2015), 1-11.

Campbell, Frances A, Craig T Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson, "Early Childhood Education: Young Adult Outcomes from the Abecedarian Project," Applied Developmental Science, 6 (2002), 42-57.

Card, David, "Chapter 30: The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, Orley C. Ashenfelter, and David Card, eds. (Elsevier, 1999).

Chandler, Michael J, "Egocentrism and Antisocial Behavior: The Assessment and Training of Social Perspective-Taking Skills," Developmental Psychology, 9 (1973), 326.

Clarke, Gregory, Hyman Hops, Peter M Lewinsohn, Judy Andrews, John R Seeley, and Julie Williams, "Cognitive-Behavioral Group Treatment of Adolescent Depression: Prediction of Outcome," Behavioral Therapy, 23 (1992), 341-354.

Cohen, Mark A, *The Costs of Crime and Justice* (Routledge, 2005).

Cohen, Mark A, Roland T Rust, Sara Steen, and Simon T Tidd, "Willingness-to-Pay for Crime Control Programs," Criminology, 42 (2004), 89-110.

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, Philip J, and Jens Ludwig, *Gun Violence: The Real Costs.* (New York: Oxford University Press, 2000).

Cunningham, Alison, "Lessons Learned from a Randomized Study of Multisystemic Therapy in Canada," in *One Step Forward*, (PRAXIS: Research from the Centre for Children & Families in the Justice System, 2002).

Currie, Janet, and Duncan Thomas, "Does Head Start Make a Difference?," American Economic Review, 85 (1995), 341-364.

Cutler, David M, and Adriana Lleras-Muney, "Chapter 2: Education and Health: Evaluating Theories and Evidence," in *Making Americans Healthier: Social and Economic Policy as Health Policy*, Robert F Schoeni, James S House, George A Kaplan, and Harold Pollack, eds. (New York: Russell Sage Foundation, 2008).

Deming, David, "Early Childhood Intervention and Life-Cycle Skill Development," American Economic Journal: Applied Economics, 1 (2009), 111-134.

Drake, Elizabeth K, Steve Aos, and Marna G Miller, "Evidence-Based Public Policy Options to Reduce Crime and Criminal Justice Costs: Implications in Washington State," Victims and Offenders, 4 (2009), 170-196.

Durlak, Joseph A, Roger P Weissberg, Allison B Dymnicki, Rebecca D Taylor, and Kriston 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, Mark, Philip Gleason, Anu Rangarajan, Robert G Wood, and Audrey Pendleton, *Impacts of Dropout Prevention Programs: Final Report* (Mathematica Policy Research, Incorporated, 1998).

Farrell, Albert D, Aleta L Meyer, Terri N Sullivan, and Eva M 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, Albert D, Aleta L Meyer, and Kamila S 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 Psychology, 30 (2001), 451-463.

Farrington, David P, and Anthony Petrosino, "The Campbell Collaboration Crime and Justice Group," The Annals of the American Academy of Political and Social Science, 578 (2001), 35-49.

Gaab, Jens, N Blättler, T Menzi, B Pabst, S Stoyer, and Ulrike Ehlert, "Randomized Controlled Evaluation of the Effects of Cognitive–Behavioral Stress Management on Cortisol Responses to Acute Stress in Healthy Subjects," Psychoneuroendocrinology, 28 (2003), 767-779.

Garces, Eliana, Duncan Thomas, and Janet Currie, "Longer-Term Effects of Head Start," American Economic Review, 92 (2002), 999-1012.

Goldin, Claudia Dale, and Lawrence F Katz, *The Race between Education and Technology* (Harvard University Press, 2009).

Governor's Crime Commission, "Juvenile Age Study: A Study of the Impact of the Jurisdiction of the Department of Juvenile Justice and Delinquency Prevention," in *Final Report to the Governor of North Carolina*, (2009).

Greenwood, Peter W, "Prevention and Intervention Programs for Juvenile Offenders," The Future of Children, 18 (2008), 185-210.

Greenwood, Peter W, and Susan Turner Rand, "Evaluation of the Paint Creek Youth Center: A Residential Program for Serious Delinquents," Criminology, 31 (1993), 263-279.

Guerra, Nancy G, and Ronald G Slaby, "Cognitive Mediators of Aggression in Adolescent Offenders: Ii. Intervention," Developmental Psychology, 26 (1990), 269.

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, Lance J Lochner, and Petra E Todd, "Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," in *Handbook of the Economics of Education*, E. Hanushek, and F. Welch, eds. (Elsevier, 2006).

Heckman, James J, Rodrigo Pinto, and Peter A Savelyev, "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes," (Cambridge, MA: National Bureau of Economic Research, Working Paper No. 18581, 2012).

Heller, Sara, Harold A Pollack, Roseanna Ander, and Jens Ludwig, "Preventing Youth Violence and Dropout: A Randomized Field Experiment," (Cambridge, MA: National Bureau of Economic Research, Working Paper No. 19014, 2013).

Henggeler, Scott W, W Glenn Clingempeel, Michael J Brondino, and Susan G Pickrel, "Four-Year Follow-up of Multisystemic Therapy with Substance-Abusing and Substance-Dependent Juvenile Offenders," Journal of the American Academy of Child & Adolescent Psychiatry, 41 (2002), 868-874.

Henggeler, Scott W, Susan G Pickrel, and Michael J Brondino, "Multisystemic Treatment of Substance-Abusing and-Dependent Delinquents: Outcomes, Treatment Fidelity, and Transportability," Mental Health Services Research, 1 (1999), 171-184.

Hudley, Cynthia, and Sandra Graham, "An Attributional Intervention to Reduce Peer-Directed Aggression among African-American 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, ed. (2011).

Illinois Juvenile Justice Commission, "Youth Reentry Improvement Report," (2011).

Illinois State Board of Education, "City of Chicago Sd 299 Per Student Spending," (https://illinoisreportcard.com/District.aspx?source=Environment&source2=PerStudentSpending&Districtid=15016299025: Illinois Report Card, 2014-2015, 2015).

In-Albon, Tina, and Silvia Schneider, "Psychotherapy of Childhood Anxiety Disorders: A Meta-Analysis," Psychotherapy and Psychosomatics, 76 (2007), 15-24.

Julian, Tiffany, and Robert Kominski, "Education and Synthetic Work-Life Earnings Estimates," in *American Community Survey Reports*, (U.S. Census Bureau, 2011).

Katz, Lawrence F, Jeffrey R Kling, and Jeffrey B Liebman, "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment," Quarterly Journal of Economics, (2001), 607-654.

Kazdin, Alan E, *Conduct Disorders in Childhood and Adolescence* (Thousand Oaks, CA: Sage Publications, 1995).

Kendall, Philip C, Mark Reber, Susan McLeer, James Epps, and Kevin R Ronan, "Cognitive-Behavioral Treatment of Conduct-Disordered Children," Cognitive Therapy and Research, 14 (1990), 279-297.

Kendall, Philip C, and Lance 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.

Klein, Nanci C, James F Alexander, and Bruce 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, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz, "Experimental Analysis of Neighborhood Effects," Econometrica, 75 (2007), 83-119.

Kling, Jeffrey R, Jens Ludwig, and Lawrence F Katz, "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment," The Quarterly Journal of Economics, (2005), 87-130.

Koegl, Christopher J, David P Farrington, Leena K Augimeri, and David 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," Psychiatric Clinics of North America, 23 (2000), 151-167.

Laird, Molly, and Steven Black, *Service-Learning Evaluation Project: Program Effects for at-Risk Students* (San Francisco, CA: Quest International, 1999).

Landenberger, Nana A, and Mark W 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, Katherine A, and Russell W Rumberger, "A.L.A.S.: Achievement for Latinos through Academic Success," in *Staying in School. A Technical Report of Three Dropout Prevention Projects for Junior High School Students with Learning and Emotional Disabilities*, H Thorton, ed. (Minneapolis, MN: University of Minnesota, Institute on Community Integration, 1995).

Lee, Stephanie, Steve Aos, Elizabeth Drake, Annie Pennucci, Marna Miller, and Laurie Anderson, "Return on Investment: Evidence Based Options to Improve Statewide Outcomes," (Olympia, WA: Washington State Institute for Public Policy, 2012).

Linehan, Marsha M, Henry Schmidt, Linda A Dimeff, J Christopher Craft, Jonathan Kanter, and Katherine A Comtois, "Dialectical Behavior Therapy for Patients with Borderline Personality Disorder and Drug-Dependence," The American Journal on Addictions, 8 (1999), 279-292.

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).

Lipsey, Mark W, Nana A Landenberger, and Sandra Jo Wilson, "Effects of Cognitive-Behavioral Programs for Criminal Offenders: A Systematic Review," in *Campbell systematic reviews*, (Nasheville, TN: Center for Evaluation Research and Methodology, Vanderbilt Institute for Public Policy Studies, 2007).

Lleras-Muney, Adriana, "The Relationship between Education and Adult Mortality in the United States," The Review of Economic Studies, 72 (2005), 189-221.

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).

Ludwig, Jens, "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, Michael S, Kurtis L Noblett, Jerry L Deffenbacher, Jackie K Gollan, and Emil F Coccaro, "Cognitive-Behavioral Therapy for Intermittent Explosive Disorder: A Pilot Randomized Clinical Trial," Journal of Consulting and Clinical Psychology, 76 (2008), 876.

McCracken, Lance M, and Dennis C Turk, "Behavioral and Cognitive-Behavioral Treatment for Chronic Pain: Outcome, Predictors of Outcome, and Treatment Process," Spine, 27 (2002), 2564-2573.

Milkman, Harvey B, and Kenneth W Wanberg, "Cognitive-Behavioral Treatment: A Review and Discussion for Corrections Professionals," (Washington, DC: US Department of Justice, National Institute of Corrections, 2007).

Miller, Ted R, Mark A Cohen, and Brian Wiersema, "Victim Costs and Consequences: A New Look," (U.S. Dept. of Justice, Office of Justice Programs, National Institute of Justice, 1996).

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

New York City Independent Budget Office, "The Rising Cost of the City's Juvenile Justice System," (2008).

Olds, David L, Charles R Henderson Jr., Harriet J Kitzman, John J Eckenrode, Robert E Cole, and Robert C 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, Rüdiger, "The Effectiveness of Social Therapy in Prison—a Randomized Experiment," Crime & Delinquency, 46 (2000), 214-232.

Page, B, and A D'Agostino, "Connect with Kids: 2004-2005: Study Results for Kansas and Missouri," (Durham, NC: Compass Consulting Group, LLC, 2005).

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," Journal of Acquired Immune Deficiency Syndromes, 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," American Journal of Public Health, 96 (2006), 1582-1587.

Pullen, Suzanne Kraus, Kim English, Bill Woodward, and Patrick C Ahlstrom, "Evaluation of the Reasoning and Rehabilitation Cognitive Skills Development Program as Implemented in Juvenile ISP in Colorado," (Office of Research and Statistics, Division of Criminal Justice, Colorado Department of Public Safety, 1996).

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, Paul, Gregory N Clarke, David E Mace, Jenel S Jorgensen, and John 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.

Rosenbaum, Michael, and Tammie Ronen, "Clinical Supervision from the Standpoint of Cognitive-Behavior Therapy," Psychotherapy: Theory, Research, Practice, Training, 35 (1998), 220.

Roush, David W, "Reforming Conditions of Confinement in Juvenile Detention: Evidence-Based Research from the U.S. District Court Intervention in Cook County, Il," Journal of Applied Juvenile Justice Services, (2015).

Sarason, Irwin G, and Victor J Ganzer, "Modeling and Group Discussion in the Rehabilitation of Juvenile Delinquents," Journal of Counseling Psychology, 20 (1973), 442-449.

Schaeffer, Cindy M, and Charles M Borduin, "Long-Term Follow-up to a Randomized Clinical Trial of Multisystemic Therapy with Serious and Violent Juvenile Offenders," Journal of Consulting and Clinical Psychology, 73 (2005), 445.

Schultz, Lynn Hickey, Dennis J Barr, and Robert L Selman, "The Value of a Developmental Approach to Evaluating Character Development Programmes: An Outcome Study of Facing History and Ourselves," Journal of Moral Education, 30 (2001), 3-27.

Schweinhart, Lawrence J, Jeanne Montie, Zongping Xiang, William S Barnett, Clive R Belfield, and Milagros Nores, "Lifetime Effects: The High/Scope Perry Preschool Study through Age 40," William S Barnett, (2005), 3.

Simons-Morton, Bruce, Denise Haynie, Keith Saylor, Aria Davis Crump, and Rusan Chen, "The Effects of the Going Places Program on Early Adolescent Substance Use and Antisocial Behavior," Prevention Science, 6 (2005), 187-197.

Skye, Dianne Lynn, "Arts-Based Guidance Intervention for Enhancement of Empathy, Locus of Control, and Prevention of Violence," (Doctoral Dissertation, University of Florida, 2001).

Steinberg, Laurence, *Age of Opportunity: Lessons from the New Science of Adolescence* (Houghton Mifflin Harcourt, 2014).

Teplin, Linda A, Karen M Abram, Gary M McClelland, Jason J Washburn, and Ann K Pikus, "Detecting Mental Disorder in Juvenile Detainees: Who Receives Services," American Journal of Public Health, 95 (2005), 1773-1780.

Timmons-Mitchell, Jane, Monica B Bender, Maureen A Kishna, and Clare C Mitchell, "An Independent Effectiveness Trial of Multisystemic Therapy with Juvenile Justice Youth," Journal of Clinical Child and Adolescent Psychology, 35 (2006), 227-236.

Toplak, Maggie E, Laura Conners, Jill Shuster, Bojana Knezevic, and Sandy 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.

Van Voorhis, Patricia, Lisa M Spruance, P Neal Ritchey, Shelley Johnson Listwan, and Renita Seabrook, "The Georgia Cognitive Skills Experiment a Replication of Reasoning and Rehabilitation," Criminal Justice and Behavior, 31 (2004), 282-305.

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.

Waldron, Holly Barrett, and Charles W Turner, "Evidence-Based Psychosocial Treatments for Adolescent Substance Abuse," Journal of Clinical Child and Adolescent Psychology, 37 (2008), 238-261.

Walker, Julian S, and Jenifer 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.

Westfall, Peter H, and Stanley S Young, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment* (New York: John Wiley & Sons, 1993).

Wood, Alison, Richard Harrington, and Anne 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, Jeffrey 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).

Zonnevylle-Bender, Marjo JS, Walter Matthys, Nicolle MH Van de Wiel, and John E Lochman, "Preventive Effects of Treatment of Disruptive Behavior Disorder in Middle Childhood on Substance Use and Delinquent Behavior," Journal of the American Academy of Child & Adolescent Psychiatry, 46 (2007), 33-39.

**Figure A.1 Study 1 and 2 Schools by Neighborhood Homicide Rate**

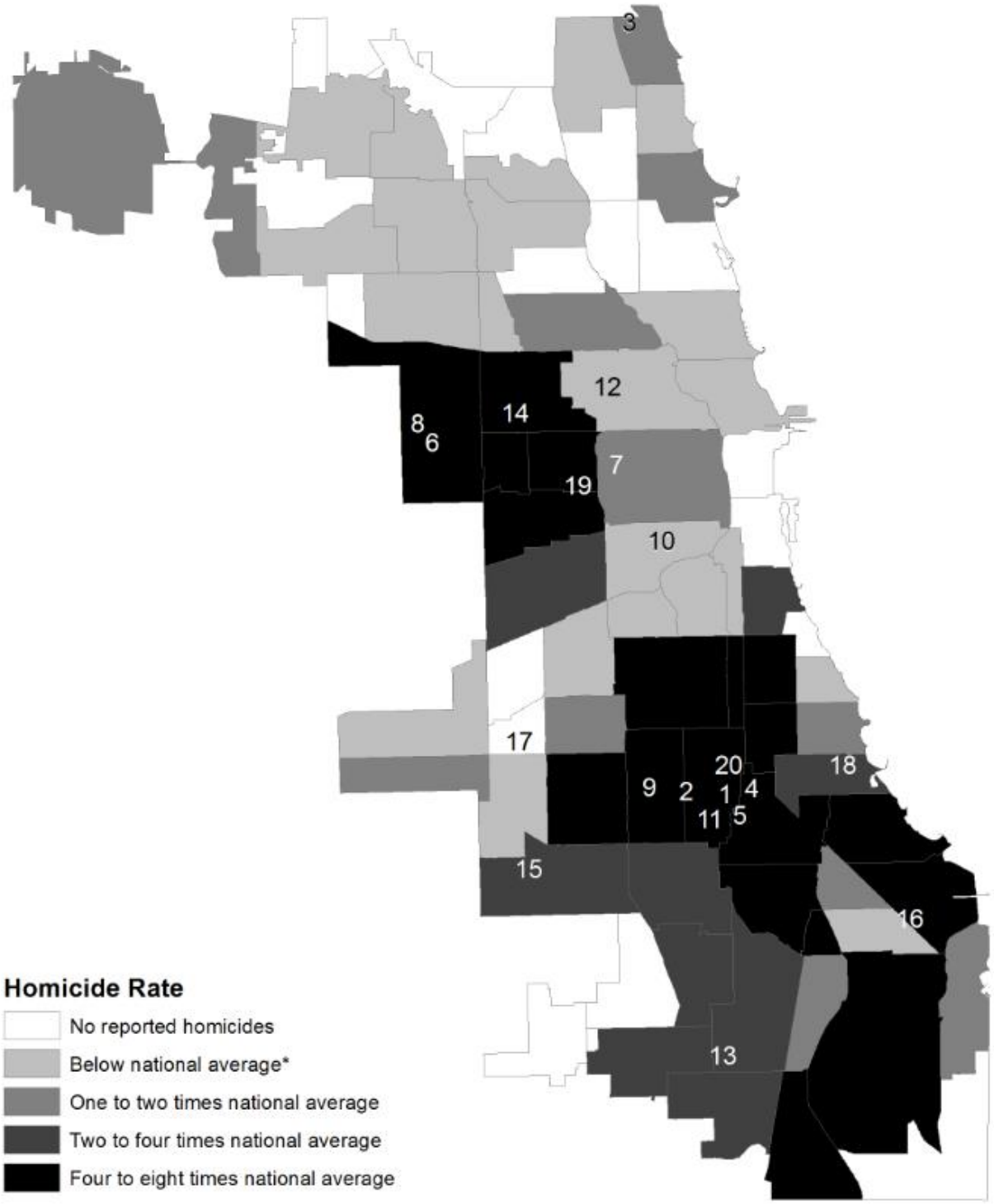


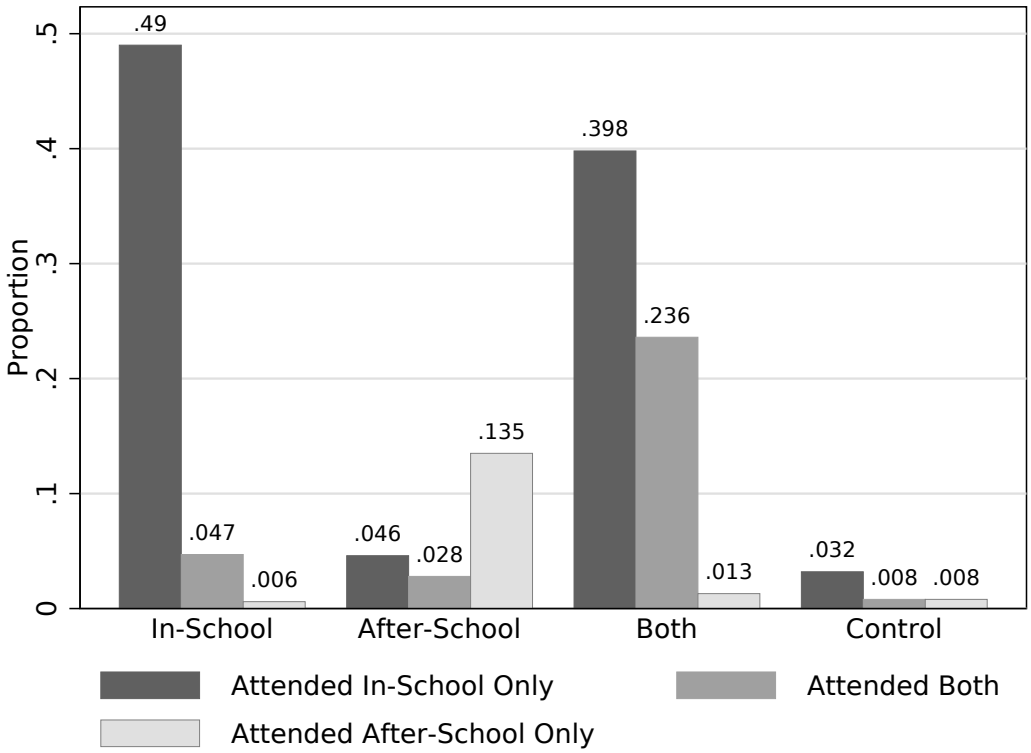
**Figure A.1 Legend**

1 Banneker ES†
2 Bass ES†
3 Jordan ES†
4 Parker Community Academy ES†
5 Yale ES†
6 Austin Polytechnic HS†
7 Crane HS†
8 Douglass HS†
9 Harper HS†
10 Juarez HS†
11 Robeson HS†
12 Clemente HS†§
13 Fenger HS†§
14 Orr HS†§
15 Bogan HS§
16 Bowen HS§
17 Hancock HS§
18 Hyde Park HS§
19 Manley HS§
20 Noble Street Charter HS – Johnson§

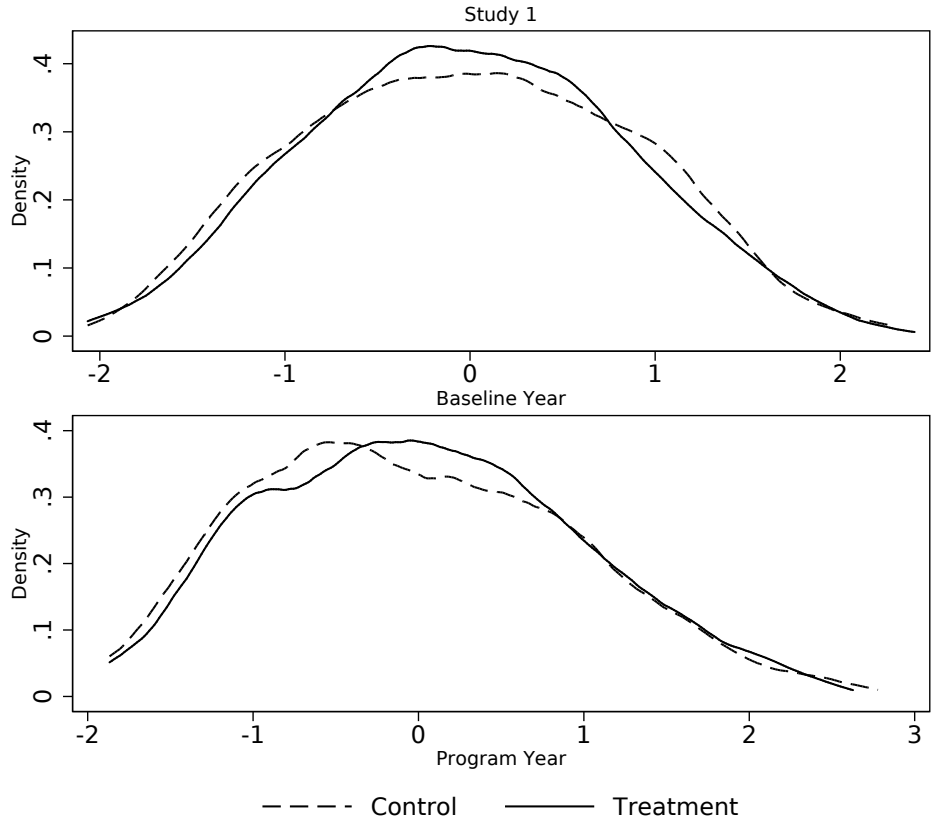
† Study 1 School
§ Study 2 School

*Rate for cities with population over 1,000,000 is 7.5 per 100,000 citizens
Sources: City of Chicago; US Census Bureau; FBI Uniform Crime Report 2013

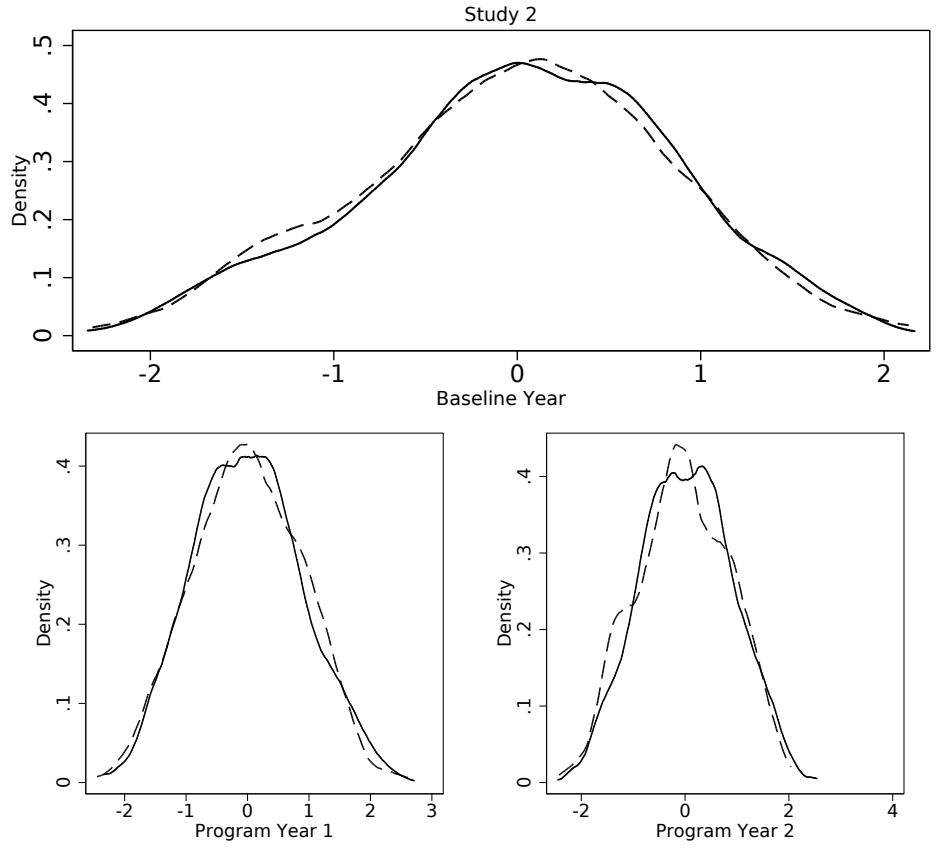
**Figure A.2 Participant Crossover, Study 1 (BAM 2009-10 Cohort)**



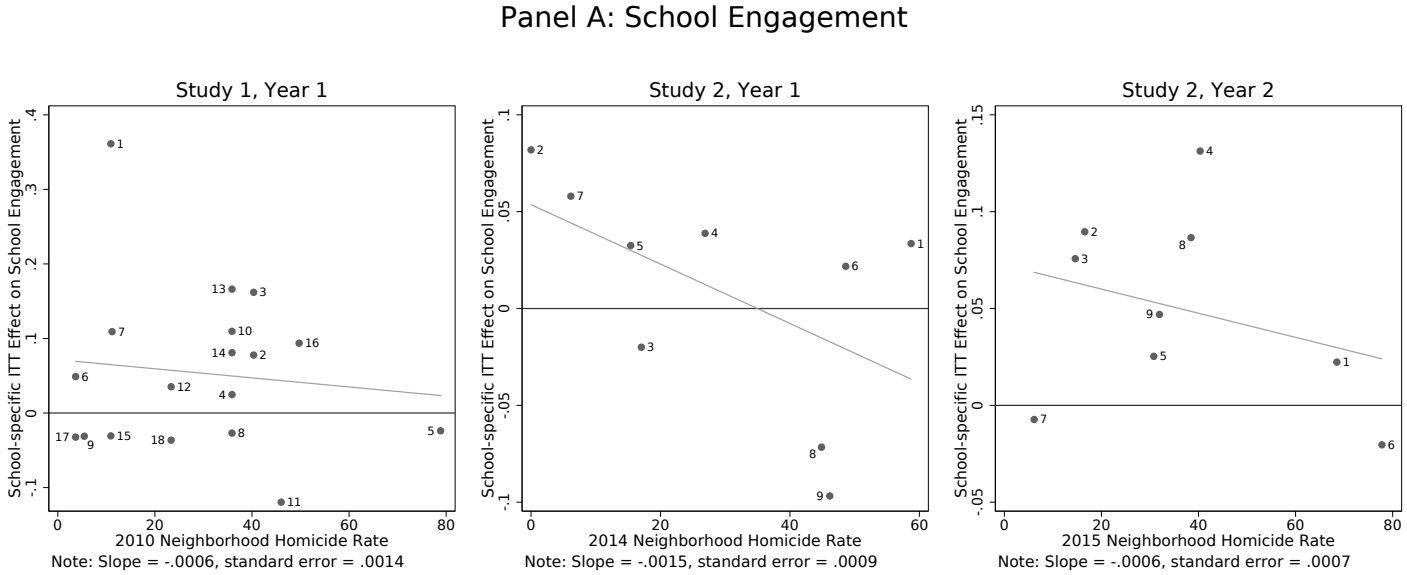
## Figure A.3 GPA Distribution Relative to School Average by Treatment

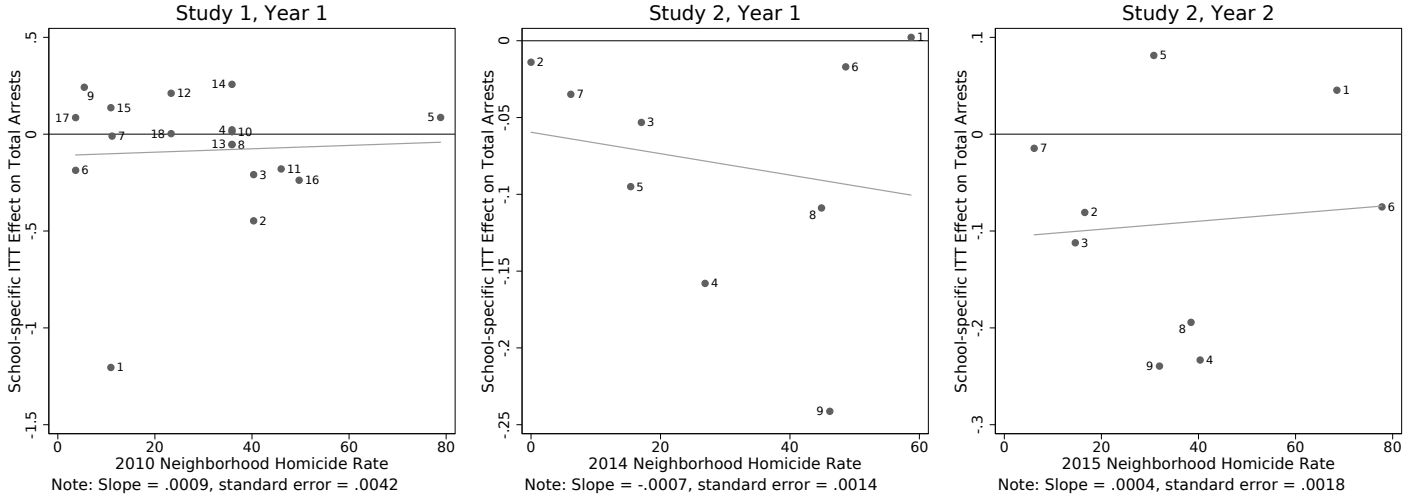



Epanechnikov kernel bandwidth .19, Adjusted for school fixed effects



# Figure A.4 School-Specific BAM Treatment Effects in Studies 1 & 2 by Local-Area Homicide Rate

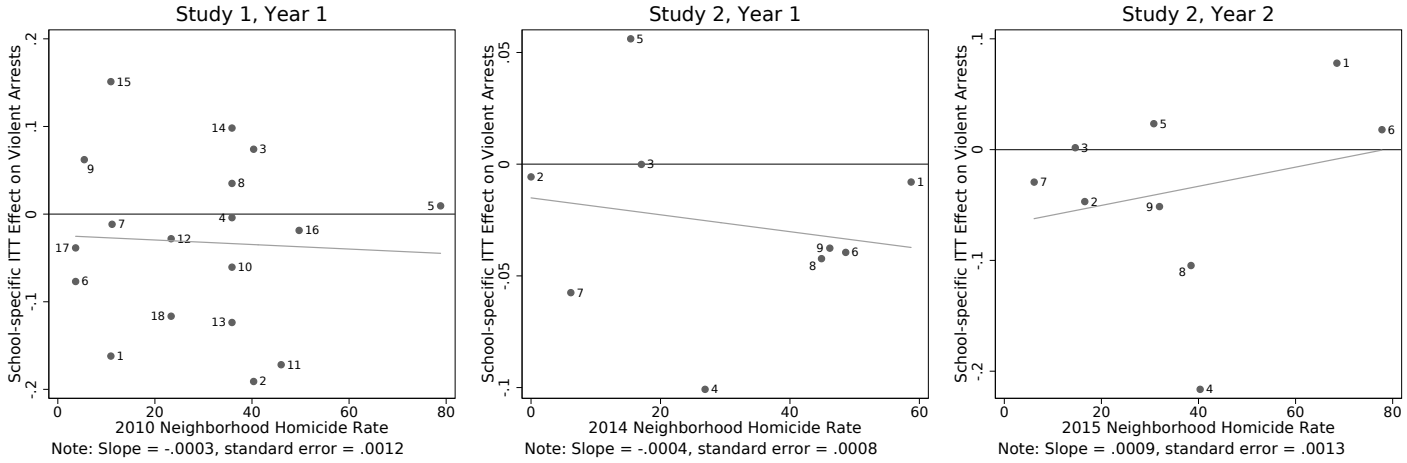



Note: Points show school-specific intention to treat effect plotted against homicide rate per 100,000 for the community area in which the school is located. School-specific effects estimated from an ITT regression that interacts treatment assignment with school fixed effects, controlling for baseline covariates and randomization block fixed effects. Line through the points in each figure fitted with OLS.

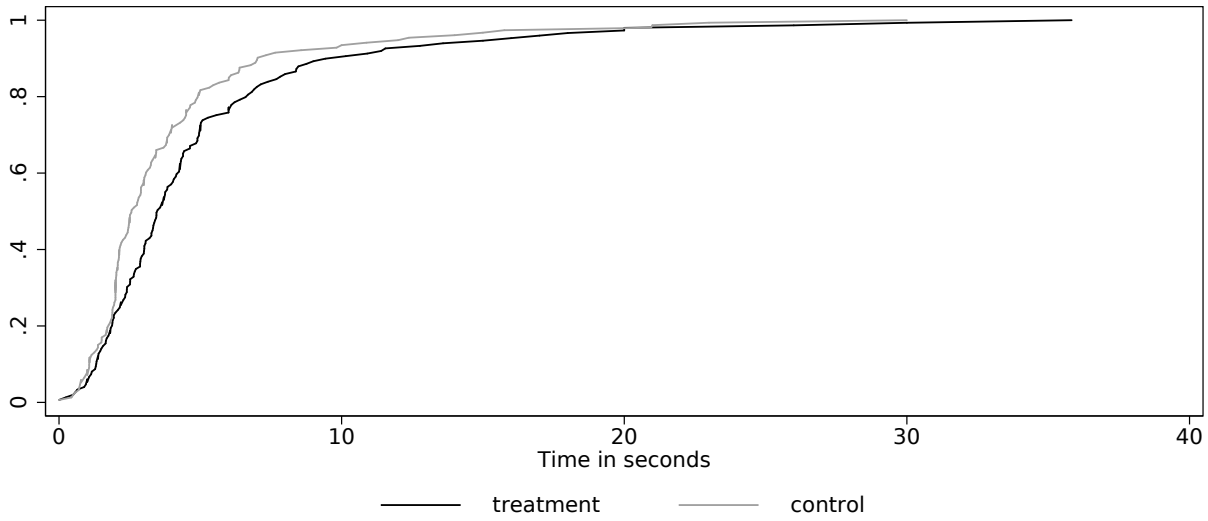
# Figure A.4 School-Specific BAM Treatment Effects in Studies 1 & 2 by Local-Area Homicide Rate (continued)

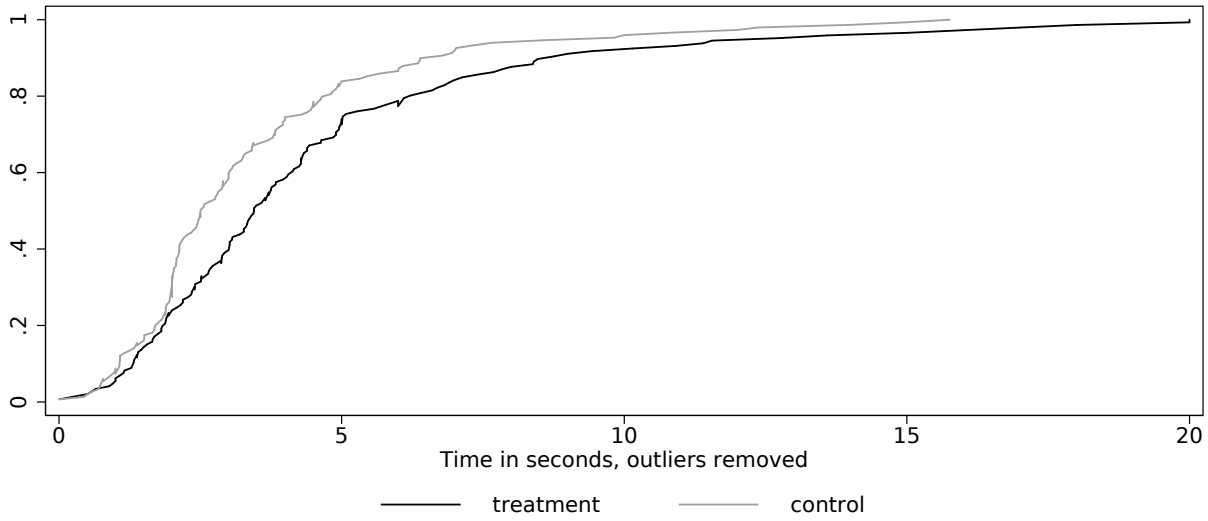
Panel C: Violent Crime Arrests

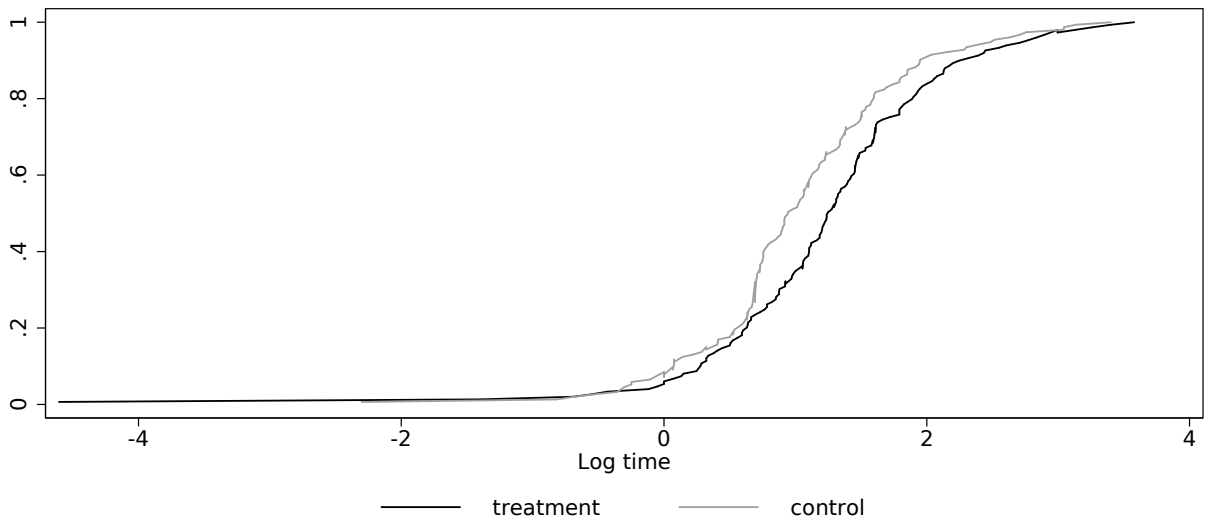



Note: Points show school-specific intention to treat effect plotted against homicide rate per 100,000 for the community area in which the school is located. School-specific effects estimated from an ITT regression that interacts treatment assignment with school fixed effects, controlling for baseline covariates and randomization block fixed effects. Line through the points in each figure fitted with OLS.

**Figure A.5 Cumulative Distribution Functions for Decision-Making Time to Iterated Dictator Game by Treatment Assignment Status for Youth in BAM Study 2**

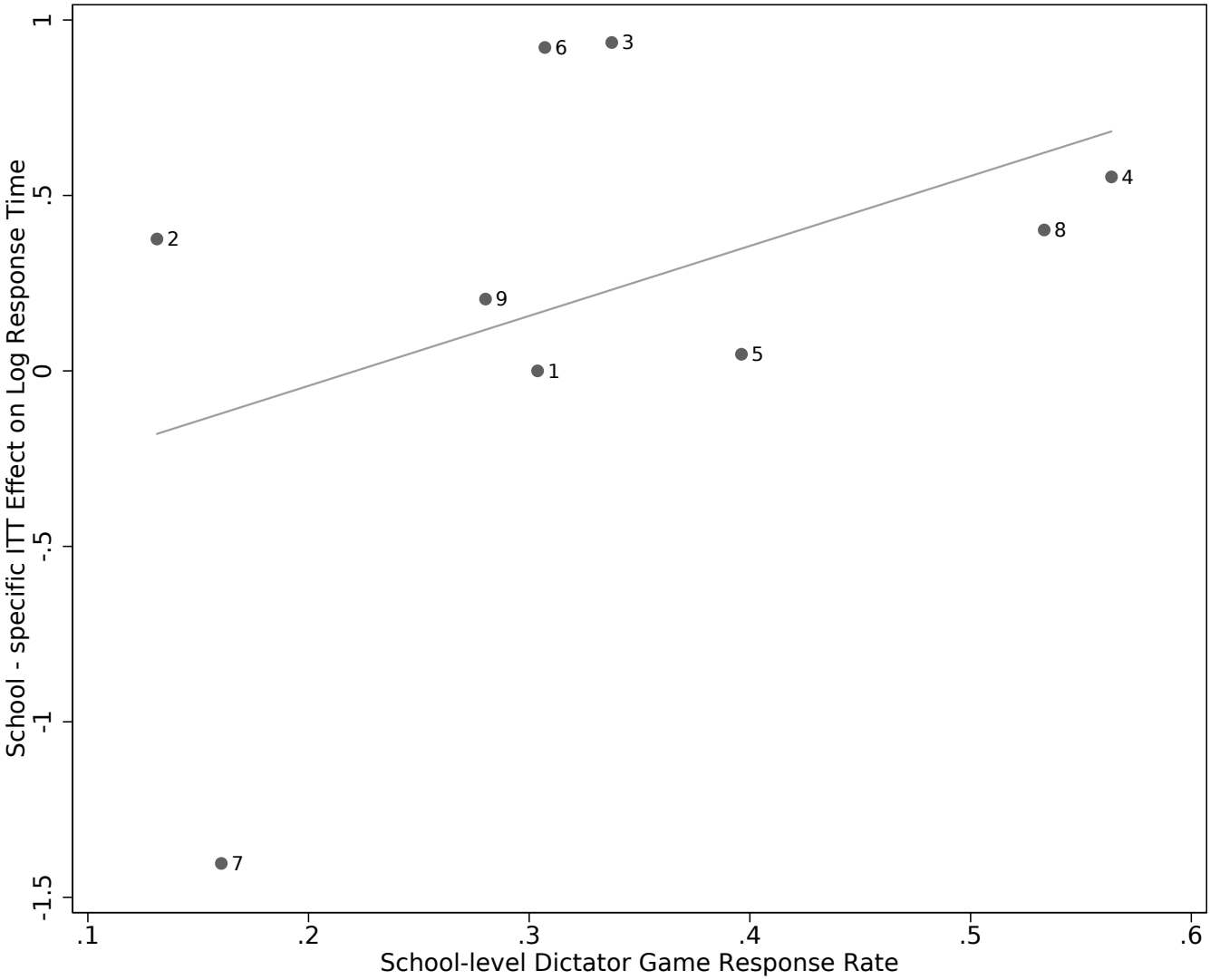



Note: Top panel shows treatment and control CDF for decision time results from administering an iterated dictator game to sub-sample of youth in BAM study 2. Middle panel shows same CDF but with youth taking longer than 20 seconds to make a decision removed. Bottom panel shows CDF for decision time after log transformation undertaken to reduce outlier influence.

## Figure A.6 School-Specific BAM Treatment Effect on Decision-Time Outcome by School-Level Reponse Rate




Note: Points show school-specific intention to treat effect plotted against school-level response rate for iterated dictator game given to sub-sample of youth in BAM study 2. School-specific effects estimated from an ITT regression that interacts treatment assignment with school fixed effects, controlling for baseline covariates and randomization block fixed effects.
Line through the points in each figure fitted with OLS.

**Table A.1. Summary of CBT Studies**

| Author(s) | Total sample size | Treatment group size | Control group size | Randomization level | Sample | Age range | T&C statistically balanced at baseline? | Limitations/concerns |
|---|---|---|---|---|---|---|---|---|
| **Farrell, Albert; Aleta Meyer; and Kamila White (2001)** | 626 | 305 | 321 | classrooms | middle school students | 10 to 15 | yes | High sample attrition; attrited students were older, had lower grade point averages, lower attendance, and more out of school suspensions |
| **Farrell, Albert; Aleta Meyer; Terri Sullivan; and Eva Kung (2003)** | 476 | 239 | 237 | classrooms | middle school students | 12 to 14 | yes | High sample attrition after one year; attrited students were older and had lower grade point averages (not significant), and were less likely to come from two-parent households (significant); self-reported outcomes |
| **Gundersen, Knut; and Frode Svartdal (2006)** | 65 | 47 | 18 | 11 sub-studies; 5 sub-studies not randomized | children and adolescents with behavior problems | mean age of 14.1 for girls, 12.6 for boys | no | Subjects were divided into eleven sub-studies, each of which was supposed to block randomize its subjects on the basis of comparable behavioral problems. In five of these sub-studies, randomization was not possible |
| **Harrington, Nancy; Steven Giles; Rick Hoyle; Greg Feeney; and Stephen Yungbluth (2001)** | 1655 | 629 (specialist condition) / 287 (teacher condition) | 739 | matched pairs of schools based on demographics and free/reduced lunch | middle school students | 11 to 13 | no | Self-reported outcomes; high sample attrition; inconsistent attendance data prevented measurement of treatment dosage; format, scale, outcome measures for violence unclear |
| **Laird, Molly; and Steven Black (1999)** | 61 | 26 | 35 | no randomization | high school students | 14 to 18 | no | No randomization; treatment and control not balanced at baseline; small sample size; only program completers considered for evaluation |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Orpinas, Pamela; Steve Kelder; Ralph Frankowski; Nancy Murray; Qing Zhang; and Alfred McAlister (2000)** | 2246 | 1020 | 1226 | schools | middle school students | 10 to 14 | yes | Treatment dosage unclear; high sample attrition; attrited students more likely to exhibit aggressive behavior; self-reported outcomes |
| **Page, B.; and A. D'Agostino (2005)** | >800 students, 46 class-rooms in 12 schools | 24 schools | 22 schools | quasi-experimental | elementary, middle, and high school students | 5 to 18 | no | Quasi-experimental; treatment dosage varied widely; treatment and control not shown to be balanced at baseline |
| **Dynarski, Mark; Philip Gleason, Anu Rangarajan, Robert Wood; and Audrey Pendleton (1998)** | 494 (cohorts 1 and 2) | 259 (cohorts 1 and 2) | 235 (cohorts 1 and 2) | individuals | students in nine CA middle schools. Mixture of students at high- and lower academic risk | mean age 14. Late middle school/ high school | yes | Slightly differential response rate in treatment and control (92% vs 86%). Randomized evaluation on subset of 219 students |
| **Patton, George; Lyndal Bond; John Carlin; Lyndal Thomas; Helen Butler; Sara Glover; Richard Catalano; and Glenn Bowes (2006)** | 26 schools | 12 schools | 14 schools | school districts | 8th grade students | 13 to 14 | no formal hypothesis tests, “little difference at baseline in key outcome measures” | Schools not shown to be balanced at baseline; 6 schools dropped out after being selected and were not included in analysis; 1 school stopped participating during intervention and was excluded from final analysis. Details of "social and emotional skills" treatment unclear; self-reported outcomes; between 19-34% of students in schools not surveyed |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Schultz, Lynn; Dennis Barr; and Robert Selman (2001)** | 346 | 212 | 134 | no randomization; four experienced FHAO teachers were used for the treatment group, and control classrooms were selected from different schools in their communities | 8th graders | 12 to 14 | no | No randomization; self-reported outcomes; treatment and control not balanced at baseline; treatment dosage unclear |
| **Simons-Morton, Bruce; Denise Haynie; Keith Saylor; Aria Davis Crump; and Rusan Chen (2005)** | 1465 | 773 | 692 | schools | 6th to 9th graders | 11 to 15 | no | Schools not shown to be balanced at baseline; high attrition; self-reported outcomes; attrited students more likely to be African-American, exhibit antisocial behavior, and come from single-parent households; attrition higher among African-Americans in treatment group than in control |
| **Larson, Katherine; and Rumberger, Russell (1995)** | 94 | 46 | 48 | individuals | mostly learning disabled/ severely emotionally disabled sample of Latino youth in one CA junior high school | 7th grade cohort | yes | Data available for students who remained in a district school, low statistical power |
| **Skye, Diane (2001)** | 153 | 78 | 75 | classrooms | high school students | 14 to 18 | no | Treatment and control groups not balanced at baseline; self-reported outcomes; treatment dosage unclear |
| **Alexander, James; and Bruce Parsons (1973)** | 86 (initial sample of 99; attrited | 46 (FFT); 30 families in comparison treatments | 10 | families | families referred through family court | 13 to 16 | yes | Imperfect randomization, based on program availability; no reporting on attrited families; little information provided on sample |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | families not reported) | | | | | | | characteristics or treatment characteristics; small sample size |
| **Armstrong, Todd (2003)** | 256 | 129 | 127 | individuals | male residents of Montgomery County Detention Center | 15 to 22 | yes, except imbalanced on racial variables | Crossover concerns; 14.7% of treatment group did not receive treatment, 19.7% of control group received treatment due to disciplinary infractions, language barriers, or residence in the wrong unit of the facility. These individuals excluded from the analysis (n=212) |
| **Borduin, Charles; Barton Mann; Lynn Coe; Scott Henggeler; Bethany Fucci; David Blaske; and Robert Williams (1995)** | 176 | 92 | 84 | families | juvenile offenders (67.5% male) | 12 to 17 | yes | Small sample. MST more comprehensive than CBT and includes family interventions. Characteristics of efficacy trial |
| **Chandler, Michael (1973)** | 45 | 15 (experimental filmmaking program) / 15 (film workshop) | 15 | individuals | delinquent boys | 11 to 13 | no | Small sample size; treatment and control not balanced at baseline; high attrition |
| **Cunningham, Alison (2002)** | 409 | "about 200" | "about 200" | families | serious youth offenders and their families (74% male) | mean 14.6; 6.6% under 12 | no | MST more comprehensive than CBT and includes family interventions; limited outcome data, as sample attrition increases rapidly after 1 year. No baseline comparison of treatment and control groups |
| **Greenwood, Peter; and Susan Turner (1993)** | 150 | 75 | 75 | individuals | male youth convicted of serious felonies | 15 to 18 | yes | 23% of treatment group removed for disciplinary reasons; unable to separate effect of CBT from other treatment components; 16% attrition rate; self-reported outcomes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Guerra, Nancy; and Ronald Slaby (1990)** | 120 | 40 (CBT) and 40 (attention control) | 40 | block randomization by gender (followed by random elimination of individuals to equalize group sizes) | male and female (equally divided) youth convicted of at least one violent crime | 15 to 18 | not specifically indicated | Small sample size; high sample attrition rate |
| **Pullen, Suzanne (1996)** | 40 | 20 | 20 | block randomization by county jurisdiction | juveniles sentenced to Juvenile Intensive Surveillance Program | not reported (mean age 16.4 years) | no | Small sample size; no baseline comparison between treatment and control; staff could make exceptions to randomized assignment; control group contained twice as many violent offenders as treatment group (40% vs. 20%); program was "barely implemented…information was imparted but skills were not developed." |
| **Henggeler, Scott; W. Glenn Clingempeel; Michael Brondino; and Susan Pickrel (1999; 2002)** | 118 | 59 | 59 | families | juvenile offenders meeting DSM-III criteria for substance abuse of dependence and their families | mean 15.7 | significant between-group baseline differences in self-reported alcohol/marijuana use and self-reported other drug use | Small sample; significant sample attrition. MST more comprehensive than CBT and includes family interventions |
| **Klein, Nanci; James Alexander; and Bruce Parsons (1977)** | 86 (initial sample of 99; attrited families not reported) | 46 (FFT); 30 families in comparison treatments | 11 | families | families referred through family court | 13 to 16 | no formal hypothesis tests | Imperfect randomization based on program availability; no reporting on attrited families; little information provided on sample characteristics or treatment characteristics; small sample size |
| **Sarason, Irwin; and Victor** | 192 | 64 (social modeling) / | 64 | individuals | male juvenile offenders | 15.5 to 18 | balanced on age, IQ, and | Sample only "essentially" randomized; integrity of |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Ganzer (1973)** | | 64 (discussion-based) | | | | | type/ severity of delinquent behavior | randomization varied by availability |
| **Schaeffer, Cindy; and Charles Borduin (2005)** | 176 | 92 | 84 | families | juvenile offenders and their families | 12 to 17 | yes | MST more comprehensive than CBT. Small sample, characteristics of efficacy trial |
| **Timmons-Mitchell, Jane; Monica Bender; Maureen Kishna; and Clare Mitchell (2006)** | 93 | 48 | 45 | families | youth who appeared before a county family court | mean 15.1 | yes | Small sample size; MST more comprehensive than CBT and includes family interventions |
| **Zonnevylle-Bender, Marjo; Walter Matthys; Nicolle van de Wiel; and John Lochman (2007)** | 77 | 38 | 39 | individuals | children exhibiting Disruptive Behavior Disorder entering 4 psychiatric outpatient clinics and 3 mental health centers; 90% male | 8 to 13 | yes | Small sample; 20.8% sample attrition; self-reported outcome |

**Table A.2. Summary of Limitations of Prior Studies of CBT**

| Author(s) | Randomization limitations | Small sample (treatment group < 100) | Data collection concerns | Attrition concerns | Self-reported behaviors as outcomes | Treatment dosage unclear /idiosyncratic |
|---|---|---|---|---|---|---|
| Laird & Black (1999) | X | X | X | | | |
| Alexander & Parsons (1973) | X | X | X | | | |
| Klein, Alexander & Parsons (1977) | X | X | X | | | |
| Chandler (1973) | X | X | | X | | |
| Guerra & Slaby (1990) | X | X | | X | | |
| Henggeler, Clingempeel, Brondino & Pickrel (2002) | X | X | | X | | |
| Skye (2001) | X | X | | | X | X |
| Pullen (1996) | X | X | | | | X |
| Gundersen & Svartdal (2006) | X | X | | | | |
| Harrington, Giles, Hoyle, Feeney & Yungbluth (2001) | X | | X | X | X | X |
| Patton, Bond, Carlin, Thomas, Butler, Glover, Catalano & Bowes (2006) | X | | X | X | X | X |
| Cunningham (2002) | X | | X | X | | |
| Simons-Morton, Haynie, Saylor, Davis Crump & Chen (2005) | X | | | X | X | |
| Schultz, Barr & Selman (2001) | X | | | | X | X |
| Page & D'Agostino (2005) | X | | | | | X |
| Sarason & Ganzer (1973) | X | | | | | |
| Armstrong (2003) | X | | | | | |
| Greenwood & Turner (1993) | | X | | | X | |
| Larson & Rumberger (1995) | | X | X | | | |
| Zonnevylle-Bender, Matthys, van de Wiel & Lochman (2007) | | X | | X | X | |
| Schaeffer & Borduin (2005) | | X | | | | |
| Borduin, Mann, Coe, Henggeler, Fucci, Blaske & Williams (1995) | | X | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Timmons-Mitchell, Bender, Kishna & Mitchell (2006) | X | | | | | |
| Dynarski, Gleason, Rangarajan, Wood, Pendleton (1998) | | X | | | | |
| Orpinas, Kelder, Frankowski, Murray, Zhang & McAlister (2000) | | | X | X | X | |
| Farrell, Meyer, Sullivan & Kung (2003) | | | X | X | | |
| Farrell, Meyer & White (2001) | | | X | | | |

**Randomization limitations**: No randomization; treatment/control not balanced at baseline; failure to report balance

**Small sample**: Treatment group <100 individuals

**Data collection concerns**: Data only collected from individuals who completed intervention; failure to collect individual data; low overall response rate

**Attrition concerns**: Attrition at least 20% or imbalanced attrition between treatment and control groups

**Self-reported behavior as outcomes**: Youths' self-reported outcomes on sensitive matters used as dependent variables

**Treatment dosage unclear / idiosyncratic**: Intervention content or intensity unclear, or heterogeneous treatments provided to different members of the treatment group

**Table A.3 Becoming a Man Studies 1 and 2 - Program Participation**

| | Panel A: BAM Study 1 (Program Year 2009-10) | | | | |
|---|---|---|---|---|---|
| | All Treatment | In-School Only | In- & After-School | After-School Only | Control |
| Ever Attend | 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 | 18 | 22 | 11 | 18 |
| | **Panel B: BAM Study 2 (Program Years 2013-14 & 2014-15)** | | | | |
| | Treatment Year 1 | Treatment Year 2 | Treatment, Both Years | Control Year 1 | Control Year 2 |
| Ever Attend | 0.51 | 0.31 | 0.52 | 0.01 | 0.02 |
| Total Sessions Attended | 8.61 | 6.71 | 15.32 | 0.22 | 0.47 |
| Total Sessions \| Ever Attended | 16.79 | 21.07 | 29.08 | 14.00 | 19.70 |
| 25th Percentile of Attenders | 8 | 12 | 11 | 6 | 7 |
| 75th Percentile of Attenders | 22 | 28 | 42 | 16 | 30 |

Notes: Panel A shows session attendance by treatment subgroup in Study 1, panel B shows session attendance by program year in Study 2. Both panels combine BAM and sports participation. Total Sessions | Ever Attended is the mean number of sessions attended by students who participated in a program activity at least once. Study 1 n = 2,740, Study 2 n = 2,064.

**Table A.4 Becoming a Man Study 1 – Program Effects on Youth Outcomes**

| | Control Mean | Intention to Treat | Effect of Participation (IV) | Lower-bound Effect of Participation (IV) | Control Complier Mean |
|---|---|---|---|---|---|
| | **Panel A: BAM Study 1, Year 1 (program offered)** | | | | |
| School Engagement Index | 0 | 0.0569*** | 0.1367*** | 0.0877*** | .222 |
| | | (0.0215) | (0.0511) | (0.0328) | |
| *Arrests Per Youth Per Year:* | | | | | |
| All Offenses | 0.699 | -0.0778* | -0.1869* | -0.1199* | .672 |
| | | (0.0456) | (0.1087) | (0.0697) | |
| Violent Offenses | 0.167 | -0.0345** | -0.0829** | -0.0532** | .186 |
| | | (0.0165) | (0.0394) | (0.0252) | |
| Property Offenses | 0.077 | 0.0048 | 0.0116 | 0.0074 | .066 |
| | | (0.0127) | (0.0303) | (0.0194) | |
| Drug Offenses | 0.151 | 0.0013 | 0.0032 | 0.0021 | .097 |
| | | (0.0177) | (0.0422) | (0.0271) | |
| Other Offenses | 0.305 | -0.0495* | -0.1188* | -0.0762* | .323 |
| | | (0.0272) | (0.0648) | (0.0415) | |
| | **Panel B: BAM Study 1, Year 2 (program not offered)** | | | | |
| School Engagement Index | 0 | 0.0782*** | 0.1878*** | 0.1206*** | .040 |
| | | (0.0215) | (0.0514) | (0.0329) | |
| *Arrests Per Youth Per Year:* | | | | | |
| All Offenses | 0.595 | -0.0643 | -0.1543 | -0.0990 | .606 |
| | | (0.0420) | (0.1000) | (0.0641) | |
| Violent Offenses | 0.11 | 0.0006 | 0.0013 | 0.0009 | .092 |
| | | (0.0143) | (0.0340) | (0.0218) | |
| Property Offenses | 0.057 | -0.0034 | -0.0082 | -0.0052 | .052 |
| | | (0.0103) | (0.0245) | (0.0157) | |
| Drug Offenses | 0.164 | -0.0196 | -0.0471 | -0.0302 | .173 |
| | | (0.0194) | (0.0461) | (0.0296) | |
| Other Offenses | 0.264 | -0.0418 | -0.1004 | -0.0644 | .288 |
| | | (0.0259) | (0.0617) | (0.0396) | |

Notes: n = 2,740. Baseline covariates and randomization block fixed effects included in all mdoels. Heteroskedasticity-robust standard errors in parentheses. Lower bound uses LATE estimates adjusted for attendance under-reporting. CCM based on main LATE estimate. * p<0.1, **p<0.05, *** p<0.01.

**Table A.5 Becoming a Man Pooled Studies 1 and 2 – Effects on Youth Outcomes**

| | Pooled Program Effects | | | | $H_0$: Program Effect = 0 | | | | | | | | $H_0$: Study 1 Effect = Study 2 Effect |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | FDR One-Stage q value | | FDR Two-Stage q value | | FWER p-value | | |
| | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Unadjusted p-value | Permuted p-value | Family = Schooling Index and 4 Crime Categories | Family = 4 Crime Categories and Total Crime | Family = Schooling Index and 4 Crime Categories | Family = 4 Crime Categories and Total Crime | Family = Schooling Index and 4 Crime Categories | Family = 4 Crime Categories and Total Crime | Unadjusted p-value |
| School Engagement | 0 | 0.0398** | 0.0880*** | 0.203 | 0.010 | .018 | 0.028 | 0.011 | 0.029 | 0.011 | 0.055 | 0.010 | 0.358 |
| | | (0.0155) | (0.0338) | | | | | | | | | | |
| Total arrests per youth per year | 0.603 | -0.0727** | -0.1611** | 0.601 | 0.019 | .019 | - | 0.032 | - | 0.034 | - | 0.054 | 0.659 |
| | | (0.0310) | (0.0683) | | | | | | | | | | |
| Violent | 0.136 | -0.0269** | -0.0597** | 0.148 | 0.013 | .012 | 0.028 | 0.032 | 0.029 | 0.034 | 0.055 | 0.054 | 0.823 |
| | | (0.0109) | (0.0239) | | | | | | | | | | |
| Property | 0.069 | 0.0026 | 0.0058 | 0.064 | 0.751 | .749 | 0.751 | 0.751 | 0.430 | 0.430 | 0.909 | 0.909 | 0.425 |
| | | (0.0082) | (0.0181) | | | | | | | | | | |
| Drug | 0.132 | -0.0048 | -0.0106 | 0.116 | 0.701 | .699 | 0.751 | 0.751 | 0.430 | 0.430 | 0.909 | 0.909 | 0.509 |
| | | (0.0124) | (0.0273) | | | | | | | | | | |
| Other | 0.266 | -0.0436** | -0.0966** | 0.273 | 0.016 | .016 | 0.028 | 0.032 | 0.029 | 0.034 | 0.055 | 0.054 | 0.937 |
| | | (0.0182) | (0.0400) | | | | | | | | | | |

Notes: n = 4,804 (all observations from studies 1 and 2 pooled together). Baseline covariates and randomization block fixed effects included. Standard errors (in parentheses) are clustered on individuals to account for two students who are in both studies. School engagement index is equal to an unweighted average of days present, GPA, and enrollment status at end of school year, all normalized to Z-score form using control group's distribution. The pooled variables capture the program years: year 1 for study 1 and years 1 and 2 combined for study 2. For study 2, the combined schooling index is an average of the index across the two program years, and the combined arrests are an average across the available data in years 1 and 2 (sum over 19 months of arrests / 2). To account for the different number of months covered by the arrest data, we test equality across the two studies by extrapolating the monthly rate of offending to a 12 month period. The permutation test results reported in the table above present pairwise-comparison p-values calculated using a re-randomization test. FDR one-stage q-value is calculated using the procedure from Benjamini and Hochberg (1995). The two-stage FDR q-value is calculated using the procedure from Benjamini, Krieger and Yekutieli (2006). The FWER p-value is calculated using the bootstrap re-sampling technique from Westfall and Young (1993). We calculate these values using two definitions of our 'family' of outcomes, first defining the family as our schooling variable plus the measures of arrests for different specific offense categories (violent, property, drug, other), excluding total arrests since it is a linear combination of the other four crime-type-specific measures; and then again defining two separate families of outcomes, using schooling as its own 'family' and then a separate family of all of our arrest measures. * p<0.10, ** p<0.05, *** p<0.01.

**Table A.6 Becoming a Man Studies 1 and 2 – Program Effects with Multiple Hypothesis Testing Adjustments**

| | | | | | | | One-Stage | | Two-Stage | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Control Mean | Intention to Treat | Unadjusted p-value | Permuted p-value | FWER adjusted p, Family = Schooling Index and 4 Crime Categories | FWER adjusted p, Family = 4 Crime Categories and Total Crime | FDR q value, Family = Schooling Index and 4 Crime Categories | FDR q value, Family = 4 Crime Categories and Total Crime | FDR q value, Family = Schooling Index and 4 Crime Categories | FDR q value, Family = 4 Crime Categories and Total Crime |
| **Panel A: BAM Study 1 (Program Year 2009-10)** | | | | | | | | | | |
| | Year 1 (program offered) | | | | | | | | | |
| School Engagement | 0 | 0.0569 | 0.008 | .012 | 0.044 | 0.008 | 0.008 | 0.009 | 0.043 | 0.009 |
| *Arrests Per Youth Per Year:* | | | | | | | | | | |
| Violent Offenses | 0.167 | -0.0345 | 0.037 | .037 | 0.138 | 0.148 | 0.037 | 0.147 | 0.080 | 0.173 |
| Property Offenses | 0.077 | 0.0048 | 0.705 | .706 | 0.914 | 0.914 | 0.705 | 0.882 | 0.545 | 0.545 |
| Drug Offenses | 0.151 | 0.0013 | 0.940 | .939 | 0.939 | 0.939 | 0.940 | 0.941 | 0.603 | 0.603 |
| Other Offenses | 0.305 | -0.0495 | 0.069 | .069 | 0.189 | 0.209 | 0.069 | 0.147 | 0.101 | 0.173 |
| All Offenses | 0.699 | -0.0778 | 0.088 | .089 | . | 0.219 | . | 0.147 | . | 0.173 |
| | Year 2 (program not offered) | | | | | | | | | |
| School Engagement | 0 | 0.0782 | 0.000 | .005 | 0.006 | 0.000 | 0.002 | 0.001 | 0.002 | 0.001 |
| *Arrests Per Youth Per Year:* | | | | | | | | | | |
| Violent Offenses | 0.110 | 0.0006 | 0.969 | .969 | 0.969 | 0.969 | 0.969 | 0.969 | 1 | 1 |
| Property Offenses | 0.057 | -0.0034 | 0.741 | .74 | 0.933 | 0.933 | 0.927 | 0.927 | 1 | 1 |
| Drug Offenses | 0.164 | -0.0196 | 0.311 | .312 | 0.674 | 0.674 | 0.519 | 0.519 | 0.452 | 0.459 |
| Other Offenses | 0.264 | -0.0418 | 0.107 | .106 | 0.361 | 0.375 | 0.267 | 0.315 | 0.271 | 0.459 |
| All Offenses | 0.595 | -0.0643 | 0.126 | .125 | . | 0.378 | . | 0.315 | . | 0.459 |
| **Panel B: BAM Study 2 (Program Years 2013-14 & 2014-15)** | | | | | | | | | | |
| | Year 1 (program offered) | | | | | | | | | |
| School Engagement | 0 | 0.0058 | 0.814 | .834 | 0.896 | 0.814 | 0.814 | 0.814 | 1 | 1 |
| *Arrests Per Youth Per Year:* | | | | | | | | | | |
| Violent Offenses | 0.119 | -0.0180 | 0.263 | .266 | 0.735 | 0.609 | 0.658 | 0.439 | 1 | 0.781 |
| Property Offenses | 0.073 | -0.0078 | 0.543 | .55 | 0.896 | 0.767 | 0.679 | 0.543 | 1 | 0.781 |
| Drug Offenses | 0.126 | -0.0153 | 0.511 | .519 | 0.896 | 0.767 | 0.679 | 0.543 | 1 | 0.781 |
| Other Offenses | 0.273 | -0.0394 | 0.179 | .186 | 0.654 | 0.551 | 0.658 | 0.439 | 1 | 0.781 |
| All Offenses | 0.591 | -0.0806 | 0.111 | .113 | . | 0.392 | . | 0.439 | . | 0.781 |
| | Year 2 (program offered) | | | | | | | | | |
| School Engagement | 0 | 0.0501 | 0.047 | .114 | 0.270 | 0.047 | 0.123 | 0.047 | 0.141 | 0.049 |
| *Arrests Per Youth Per Year:* | | | | | | | | | | |
| Violent Offenses | 0.079 | -0.0276 | 0.074 | .075 | 0.270 | 0.256 | 0.123 | 0.123 | 0.141 | 0.141 |
| Property Offenses | 0.046 | -0.0018 | 0.856 | .857 | 0.857 | 0.857 | 0.857 | 0.857 | 0.521 | 0.521 |
| Drug Offenses | 0.094 | -0.0147 | 0.390 | .397 | 0.636 | 0.636 | 0.488 | 0.488 | 0.243 | 0.243 |
| Other Offenses | 0.163 | -0.0400 | 0.071 | .073 | 0.270 | 0.256 | 0.123 | 0.123 | 0.141 | 0.141 |
| All Offenses | 0.383 | -0.0841 | 0.032 | .033 | . | 0.131 | . | 0.123 | . | 0.141 |

Notes: Intention-to-treat estimates and unadjusted p-values (see Table IV for model detail) are presented by year for BAM study 1 and 2 alongside adjustments for multiple hypothesis testing. Family-wise error rate and false discovery rate calculations described Table A5. Study 1 n = 2,740; study 2 n = 2,064, * p<0.10, ** p<0.05, *** p<0.01.

**Table A.7 Becoming A Man Study 1 – Program Effects on Arrests Using Lower Match Quality Threshold**

| | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean |
|---|---|---|---|---|
| | Year 1 (program offered) | | | |
| *Arrests Per Youth Per Year:* | | | | |
| All Offenses | 0.727 | -0.0852* | -0.2044* | 0.744 |
| | | (0.0470) | (0.1118) | |
| Violent Offenses | 0.171 | -0.0323* | -0.0774* | 0.196 |
| | | (0.0169) | (0.0402) | |
| Property Offenses | 0.08 | 0.0059 | 0.0142 | 0.071 |
| | | (0.0131) | (0.0312) | |
| Drug Offenses | 0.154 | -0.0002 | -0.0004 | 0.110 |
| | | (0.0181) | (0.0429) | |
| Other Offenses | 0.322 | -0.0587** | -0.1408** | 0.368 |
| | | (0.0279) | (0.0666) | |
| | Year 2 (program not offered) | | | |
| *Arrests Per Youth Per Year:* | | | | |
| All Offenses | 0.629 | -0.0734* | -0.1762* | 0.671 |
| | | (0.0434) | (0.1033) | |
| Violent Offenses | 0.119 | -0.0057 | -0.0137 | 0.113 |
| | | (0.0149) | (0.0353) | |
| Property Offenses | 0.065 | -0.0069 | -0.0166 | 0.070 |
| | | (0.0112) | (0.0267) | |
| Drug Offenses | 0.174 | -0.0234 | -0.0561 | 0.192 |
| | | (0.0200) | (0.0476) | |
| Other Offenses | 0.272 | -0.0374 | -0.0897 | 0.296 |
| | | (0.0264) | (0.0630) | |

Notes: n = 2,740. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroskedasticity-robust standard errors in parentheses. Regressions use lower threshold in the probabilistic matching process (i.e., allow for more errors/mismatches in names and dates of birth). * p<0.1, ** p<0.05, *** p<0.01.

**Table A.8 Becoming a Man Studies 1 and 2 - Program Effects on School Engagement and Performance, Standardized Units**

| | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Number of Non-Missing Observations | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Number of Non-Missing Observations |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Panel A: BAM Study 1 (Program Year 2009-10)** | | | | | | | | | |
| | Year 1 (program offered) | | | | | Year 2 (program not offered) | | | | |
| School Engagement Index | 0 | 0.0569*** | 0.1367*** | 0.222 | 2740 | 0 | 0.0782*** | 0.1878*** | 0.040 | 2740 |
| | | (0.0215) | (0.0511) | | | | (0.0215) | (0.0514) | | |
| *Index Elements* | | | | | | | | | | |
| Days Present | 0 | 0.0436 | 0.1024 | 0.413 | 2660 | 0 | 0.0475 | 0.1023 | 0.188 | 2264 |
| | | (0.0281) | (0.0653) | | | | (0.0349) | (0.0744) | | |
| GPA | 0 | 0.0571* | 0.1259* | 0.171 | 2466 | 0 | 0.1026*** | 0.2029*** | -0.028 | 1837 |
| | | (0.0305) | (0.0666) | | | | (0.0383) | (0.0752) | | |
| Still in School | 0 | 0.0503 | 0.1208 | 0.147 | 2740 | 0 | 0.0412 | 0.099 | 0.137 | 2740 |
| | | (0.0345) | (0.0820) | | | | (0.0349) | (0.0829) | | |
| | **Panel B: BAM Study 2 (Program Years 2013-14 & 2014-15)** | | | | | | | | | |
| | Year 1 (program offered) | | | | | Year 2 (program offered) | | | | |
| School Engagement Index | 0 | 0.0058 | 0.0116 | 0.212 | 2064 | 0 | 0.0501** | 0.0993** | 0.081 | 2064 |
| | | (0.0248) | (0.0484) | | | | (0.0252) | (0.0490) | | |
| *Index Elements* | | | | | | | | | | |
| Days Present | 0 | 0.0103 | 0.0190 | 0.252 | 1912 | 0 | 0.0543 | 0.0960 | 0.094 | 1717 |
| | | (0.0343) | (0.0621) | | | | (0.0409) | (0.0710) | | |
| GPA | 0 | -0.0048 | -0.0083 | 0.059 | 1508 | 0 | 0.0805* | 0.1348* | -0.090 | 1257 |
| | | (0.0386) | (0.0651) | | | | (0.0486) | (0.0797) | | |
| Still in School | 0 | 0.0253 | 0.0502 | 0.306 | 2064 | 0 | 0.0457 | 0.0906 | 0.221 | 2064 |
| | | (0.0381) | (0.0742) | | | | (0.0396) | (0.0769) | | |

Notes: All variables standarized on the control group by year, so coefficients are in standard deviation units. Index elements use only observations with non-missing data for that element. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroskedasticity-robust standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.9 Becoming a Man Studies 1 and 2 - Program Effects on School Engagement and Performance, Original Units**

| | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Number of Non-Missing Observations | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Number of Non-Missing Observations |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Panel A: BAM Study 1 (Program Year 2009-10)** | | | | | | | | | |
| | Year 1 (program offered) | | | | | Year 2 (program not offered) | | | | |
| *Index Elements* | | | | | | | | | | |
| Days Present | 104.269 | 2.1464 | 5.042 | 124.586 | 2660 | 100.162 | 2.5097 | 5.4094 | 110.113 | 2264 |
| | | (1.3856) | (3.2128) | | | | (1.8449) | (3.9318) | | |
| GPA | 1.485 | 0.0576* | 0.1271* | 1.657 | 2466 | 1.537 | 0.1040*** | 0.2057*** | 1.509 | 1837 |
| | | (0.0308) | (0.0672) | | | | (0.0388) | (0.0763) | | |
| Still in School | 0.875 | 0.0166 | 0.0399 | 0.924 | 2740 | .758 | 0.0177 | 0.0424 | 0.817 | 2740 |
| | | (0.0114) | (0.0271) | | | | (0.0149) | (0.0355) | | |
| | **Panel B: BAM Study 2 (Program Years 2013-14 & 2014-15)** | | | | | | | | | |
| | Year 1 (program offered) | | | | | Year 2 (program offered) | | | | |
| *Index Elements* | | | | | | | | | | |
| Days Present | 130.909 | 0.4959 | 0.9139 | 143.021 | 1912 | 128.332 | 2.6980 | 4.7684 | 132.988 | 1717 |
| | | (1.6502) | (2.9892) | | | | (2.0311) | (3.5239) | | |
| GPA | 1.937 | -0.0047 | -0.0081 | 1.995 | 1508 | 1.953 | 0.0782* | 0.1308* | 1.865 | 1257 |
| | | (0.0378) | (0.0638) | | | | (0.0471) | (0.0773) | | |
| Still in School | 0.633 | 0.0122 | 0.0242 | 0.780 | 2064 | 0.515 | 0.0228 | 0.0453 | 0.626 | 2064 |
| | | (0.0184) | (0.0358) | | | | (0.0198) | (0.0385) | | |

Notes: This table is identical to table A8, except that original units are used for index elements in place of Z-score units. Days present out of 180-day school year; GPA on 4-point scale; still in school is an indicator variable for having at least 1 grade in the 4th quarter in CPS records. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroskedasticity-robust standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.10 Sensitivity Analysis of Study 1 and 2 Results to Treatment of Missing Data**

| | CM | ITT | IV | CCM | CM | ITT | IV | CCM |
|---|---|---|---|---|---|---|---|---|
| | **Panel A: BAM Study 1 (Program Year 2009-10)** | | | | | | | |
| | Year 1 (program offered) | | | | Year 2 (program not offered) | | | |
| Main Results | 0.000 | 0.0569*** | 0.1367*** | 0.222 | 0.000 | 0.0782*** | 0.1878*** | 0.040 |
| | | (0.0215) | (0.0511) | | | (0.0215) | (0.0514) | |
| Listwise Deletion | 0.000 | 0.0446* | 0.0983* | 0.202 | 0.000 | 0.0680** | 0.1343** | 0.031 |
| *Year 1 n = 2466, Year 2 n = 1833* | | (0.0237) | (0.0517) | | | (0.0301) | (0.0590) | |
| Listwise Deletion, IPW | -0.020 | 0.0457* | 0.1031* | 0.191 | -0.106 | 0.0783** | 0.1691** | -0.065 |
| *Year 1 n = 2466, Year 2 n = 1833* | | (0.0243) | (0.0542) | | | (0.0359) | (0.0769) | |
| Zero Imputation | 0.000 | 0.0610** | 0.1465** | 0.261 | 0.000 | 0.0478* | 0.1147* | 0.205 |
| | | (0.0251) | (0.0594) | | | (0.0267) | (0.0635) | |
| CPS Leave Codes | 0.000 | 0.0494** | 0.1186** | 0.143 | 0.000 | 0.0676*** | 0.1622*** | 0.025 |
| | | (0.0213) | (0.0506) | | | (0.0213) | (0.0508) | |
| Multiple Imputation | 0.000 | 0.0498** | 0.1196** | 0.258 | 0.000 | 0.0548** | 0.1315** | 0.164 |
| | | (0.0232) | (0.0550) | | | (0.0268) | (0.0638) | |
| | **Panel B: BAM Study 2 (Program Years 2013-14 & 2014-15)** | | | | | | | |
| | Year 1 (program offered) | | | | Year 2 (program offered) | | | |
| Main Results | 0.000 | 0.0058 | 0.0117 | 0.221 | 0.000 | 0.0501** | 0.0993** | 0.081 |
| | | (0.0248) | (0.0488) | | | (0.0252) | (0.0490) | |
| Listwise Deletion | 0.000 | 0.0106 | 0.0184 | 0.112 | 0.000 | 0.0383 | 0.0642 | -0.012 |
| *Year 1 n = 1506, Year 2 n = 1257* | | (0.0336) | (0.0569) | | | (0.0405) | (0.0662) | |
| Listwise Deletion, IPW | -0.025 | 0 | 0 | 0.110 | -0.083 | 0.0526 | 0.0885 | -0.080 |
| *Year 1 n = 1506, Year 2 n = 1257* | | (0.0358) | (0.0612) | | | (0.0468) | (0.0769) | |
| Zero Imputation | 0.000 | 0.0097 | 0.0194 | 0.336 | 0.000 | 0.0475 | 0.0943 | 0.225 |
| | | (0.0327) | (0.0643) | | | (0.0349) | (0.0678) | |
| CPS Leave Codes | 0.000 | 0.0182 | 0.0364 | 0.223 | 0.000 | 0.0653** | 0.1295** | 0.104 |
| | | (0.0260) | (0.0510) | | | (0.0280) | (0.0545) | |
| Multiple Imputation | 0.000 | 0.0090 | 0.0181 | 0.222 | 0.000 | 0.0574 | 0.1139 | 0.105 |
| | | (0.0284) | (0.0561) | | | (0.0383) | (0.0753) | |

Notes: Main results have group mean imputed for missing elements of engagement index. Listwise deletion forms the composite with only observations missing none of the elements while listwise deletion, IPW weights observations by inverse of predicted probability of having all non-missing data based on parsimonious set of baseline covariates. Zero imputation assumes 0 GPA and 0 days present if missing. CPS leave codes imputes group means for transfers and graduates but 0s for lost, withdrawn, corrections, deceased, or no leave code. Regressions use all observations unless otherwise noted. Coefficients in standard deviation units. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroskedasticty-robust standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.11 Becoming a Man Study 1 – Intention to Treat Effects by Treatment Arm**

| | Schooling Index | Violent Crime Arrests |
|---|---|---|
| | Year 1 (program offered) | |
| In-School | 0.0567* | -0.0237 |
| | (0.0297) | (0.0231) |
| After-School | 0.0684** | -0.0466** |
| | (0.0332) | (0.0219) |
| Both | 0.0490* | -0.0365 |
| | (0.0285) | (0.0225) |
| CM | 0 | .167 |
| | Year 2 (program not offered) | |
| In-School | 0.0710** | 0.0081 |
| | (0.0301) | (0.0212) |
| After-School | 0.0899*** | -0.0055 |
| | (0.0328) | (0.0193) |
| Both | 0.0771*** | -0.0025 |
| | (0.0288) | (0.0202) |
| CM | 0 | .110 |

Notes: n = 2,740. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroskedasticity-robust standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.12 Becoming a Man Studies 1 and 2 – Main Results Including Motor Vehicle Arrests in "All Offenses"**

| | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean |
|---|---|---|---|---|---|---|---|---|
| | **Panel A: BAM Study 1 (Program Year 2009-10, n = 2,740)** | | | | | | | |
| | Year 1 (program offered) | | | | Year 2 (program not offered) | | | |
| School Engagement Index | 0 | 0.0569*** | 0.1367*** | 0.222 | 0 | 0.0782*** | 0.1878*** | 0.040 |
| | | (0.0215) | (0.0511) | | | (0.0215) | (0.0514) | |
| *Arrests Per Youth Per Year:* | | | | | | | | |
| All Offenses | 0.716 | -0.0810* | -0.1944* | 0.694 | 0.612 | -0.0674 | -0.1618 | 0.627 |
| | | (0.0462) | (0.1100) | | | (0.0424) | (0.1011) | |
| Violent Offenses | 0.167 | -0.0345** | -0.0829** | 0.186 | 0.11 | 0.0006 | 0.0013 | 0.092 |
| | | (0.0165) | (0.0394) | | | (0.0143) | (0.0340) | |
| Property Offenses | 0.077 | 0.0048 | 0.0116 | 0.066 | 0.057 | -0.0034 | -0.0082 | 0.052 |
| | | (0.0127) | (0.0303) | | | (0.0103) | (0.0245) | |
| Drug Offenses | 0.151 | 0.0013 | 0.0032 | 0.097 | 0.164 | -0.0196 | -0.0471 | 0.173 |
| | | (0.0177) | (0.0422) | | | (0.0194) | (0.0461) | |
| Other Offenses | 0.305 | -0.0495* | -0.1188* | 0.323 | 0.264 | -0.0418 | -0.1004 | 0.288 |
| | | (0.0272) | (0.0648) | | | (0.0259) | (0.0617) | |
| | **Panel B: BAM Study 2 (Program Years 2013-14 & 2014-15, n = 2,064)** | | | | | | | |
| | Year 1 (program offered) | | | | Year 2 (program offered) | | | |
| School Engagement Index | 0 | 0.0058 | 0.0117 | 0.221 | 0 | 0.0501** | 0.0993** | 0.081 |
| | | (0.0248) | (0.0488) | | | (0.0252) | (0.0490) | |
| *Arrests Per Youth Per Year:* | | | | | | | | |
| All Offenses | 0.591 | -0.0806 | -0.1614 | 0.630 | 0.383 | -0.0841** | -0.1670** | 0.471 |
| | | (0.0506) | (0.0999) | | | (0.0392) | (0.0771) | |
| Violent Offenses | 0.119 | -0.0180 | -0.0361 | 0.121 | 0.079 | -0.0276* | -0.0549* | 0.110 |
| | | (0.0161) | (0.0318) | | | (0.0155) | (0.0303) | |
| Property Offenses | 0.073 | -0.0078 | -0.0157 | 0.075 | 0.046 | -0.0018 | -0.0036 | 0.062 |
| | | (0.0129) | (0.0253) | | | (0.0101) | (0.0197) | |
| Drug Offenses | 0.126 | -0.0153 | -0.0307 | 0.168 | 0.094 | -0.0147 | -0.0292 | 0.115 |
| | | (0.0233) | (0.0459) | | | (0.0171) | (0.0335) | |
| Other Offenses | 0.273 | -0.0394 | -0.0789 | 0.266 | 0.163 | -0.0400* | -0.0793* | 0.183 |
| | | (0.0293) | (0.0579) | | | (0.0221) | (0.0434) | |

Notes: Baseline covariates and randomization fixed effects included in all models (see text). Heteroskedasticity-robust standard errors in parentheses. School engagement index is equal to an unweighted average of days present, GPA, and enrollment status at end of school year, all normalized to Z-score form using control group's distribution. Year 1 arrest data from start of program school year until start of following school year for both studies. Year 2 arrest data through following July 18th for study 1 (about 10 months) and through following March 31st (about 8 months) for study 2. Total arrests includes motor-vehicle offenses. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.13 Juvenile Detention Study 3 – Treatment Effect on Probability of Re-admission within Given Number of Months, Balanced Panel**

| Months Since Release | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| ITT without Fixed Effects | -0.0262 | -0.0399** | -0.0449** | -0.0459** | -0.0400** | -0.0371** | -0.0425** | -0.0335** | -0.0338** |
| | (0.0173) | (0.0188) | (0.0187) | (0.0185) | (0.0180) | (0.0175) | (0.0173) | (0.0171) | (0.0170) |
| ITT with Fixed Effects | -0.0351* | -0.0495** | -0.0519*** | -0.0526*** | -0.0462** | -0.0442** | -0.0500*** | -0.0418** | -0.0430** |
| | (0.0186) | (0.0200) | (0.0199) | (0.0195) | (0.0191) | (0.0187) | (0.0184) | (0.0182) | (0.0181) |
| CM | 0.324 | 0.477 | 0.566 | 0.607 | 0.636 | 0.659 | 0.678 | 0.685 | 0.689 |
| | | | | | | | | | |
| LATE without Fixed Effects | -0.1031 | -0.1569** | -0.1765** | -0.1805** | -0.1574** | -0.1457** | -0.1672** | -0.1317* | -0.1328** |
| | (0.0679) | (0.0738) | (0.0741) | (0.0732) | (0.0713) | (0.0694) | (0.0688) | (0.0675) | (0.0670) |
| LATE with Fixed Effects | -0.1342** | -0.1893*** | -0.1984*** | -0.2010*** | -0.1764** | -0.1689** | -0.1911*** | -0.1598** | -0.1643** |
| | (0.0682) | (0.0734) | (0.0735) | (0.0722) | (0.0704) | (0.0690) | (0.0683) | (0.0673) | (0.0669) |
| CCM | 0.314 | 0.531 | 0.659 | 0.717 | 0.714 | 0.745 | 0.769 | 0.762 | 0.763 |

Notes: n = 2693. Dependent variable is indicator for whether youth returned to JTDC within X months. Some individuals have multiple spells in the data, so standard errors are clustered on the individual. Probit results similar to linear probability model shown. ITT is the intention to treat effect, while LATE is the IV estimate for the effect of participation in CBT using random assignment to CBT as an instrument for participation in CBT. LATE operationalizes treatment receipt as spending more than 5% of a stay in a treatment unit. Specifications with fixed effects include day-of-admission fixed effects. CCM calculated from model without fixed effects. Baseline covariates included in all regressions. Sample is all male spells during randomization with complete 18-month follow-up data. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.14 Juvenile Detention Study 3 – Treatment Effect on Probability of Re-admission within Given Number of Months, Balanced Panel, First Spell in Randomization Only**

| Months Since Release | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| ITT without Fixed Effects | -0.0214 | -0.0278 | -0.0384* | -0.0327 | -0.0291 | -0.0347 | -0.0401* | -0.0314 | -0.0305 |
| | (0.0201) | (0.0221) | (0.0223) | (0.0220) | (0.0216) | (0.0212) | (0.0210) | (0.0208) | (0.0207) |
| ITT with Fixed Effects | -0.0238 | -0.0371 | -0.0513** | -0.0456* | -0.0386* | -0.0456** | -0.0518** | -0.0433* | -0.0443** |
| | (0.0219) | (0.0242) | (0.0239) | (0.0235) | (0.0232) | (0.0227) | (0.0225) | (0.0223) | (0.0221) |
| CM | 0.278 | 0.423 | 0.516 | 0.553 | 0.586 | 0.613 | 0.633 | 0.640 | 0.645 |
| | | | | | | | | | |
| LATE without Fixed Effects | -0.0851 | -0.1106 | -0.1529* | -0.1304 | -0.1159 | -0.1384 | -0.1599* | -0.1252 | -0.1213 |
| | (0.0793) | (0.0875) | (0.0887) | (0.0876) | (0.0859) | (0.0845) | (0.0839) | (0.0828) | (0.0822) |
| LATE with Fixed Effects | -0.0909 | -0.1417* | -0.1960** | -0.1742** | -0.1476* | -0.1744** | -0.1983** | -0.1657** | -0.1695** |
| | (0.0776) | (0.0859) | (0.0857) | (0.0839) | (0.0826) | (0.0815) | (0.0809) | (0.0799) | (0.0794) |
| CCM | 0.269 | 0.470 | 0.618 | 0.648 | 0.660 | 0.708 | 0.734 | 0.730 | 0.731 |

Notes: n = 1860. Dependent variable is indicator for whether youth returned to JTDC within X months. Probit results similar to linear probability model shown. ITT is the intention to treat effect, while LATE is the IV estimate for the effect of participation in CBT using random assignment to CBT as an instrument for participation in CBT. LATE operationalizes treatment receipt as spending more than 5% of a stay in a treatment unit. Specifications with fixed effects include day-of-admission fixed effects. CCM calculated from model without fixed effects. Baseline covariates included in all regressions. Sample is the first spell during the randomization period for each male youth, provided that spell has complete 18-month follow-up data. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.15 Juvenile Detention Study 3 – Treatment Effect on Number of Re-Admissions within Given Number of Months, Balanced Panel**

| Months Since Release | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| ITT without Fixed Effects | -0.0315 | -0.0639** | -0.0672* | -0.0965** | -0.1069** | -0.0988* | -0.1325** | -0.1333** | -0.1323** |
| | (0.0214) | (0.0311) | (0.0380) | (0.0445) | (0.0503) | (0.0560) | (0.0602) | (0.0636) | (0.0656) |
| ITT with Fixed Effects | -0.0427* | -0.0862*** | -0.0892** | -0.1240*** | -0.1400*** | -0.1383** | -0.1808*** | -0.1839*** | -0.1843*** |
| | (0.0227) | (0.0333) | (0.0406) | (0.0474) | (0.0533) | (0.0586) | (0.0634) | (0.0671) | (0.0688) |
| CM | 0.371 | 0.679 | 0.912 | 1.112 | 1.291 | 1.450 | 1.604 | 1.710 | 1.789 |
| | | | | | | | | | |
| LATE without Fixed Effects | -0.124 | -0.2511** | -0.2640* | -0.3792** | -0.4200** | -0.3882* | -0.5208** | -0.5239** | -0.5201** |
| | (0.0842) | (0.1227) | (0.1496) | (0.1753) | (0.1977) | (0.2194) | (0.2365) | (0.2498) | (0.2574) |
| LATE with Fixed Effects | -0.1633* | -0.3295*** | -0.3411** | -0.4739*** | -0.5353*** | -0.5286** | -0.6911*** | -0.7031*** | -0.7045*** |
| | (0.0835) | (0.1235) | (0.1493) | (0.1746) | (0.1956) | (0.2148) | (0.2331) | (0.2464) | (0.2522) |
| CCM | 0.348 | 0.737 | 0.964 | 1.292 | 1.463 | 1.593 | 1.825 | 1.937 | 2.011 |

Notes: n = 2693. Dependent variable is count of how many times youth returned to JTDC within X months. Some individuals have multiple spells in the data, so standard errors are clustered on the individual. ITT is the intention to treat effect, while LATE is the IV estimate for the effect of participation in CBT using random assignment to CBT as an instrument for participation in CBT. LATE operationalizes treatment receipt as spending more than 5% of a stay in a treatment unit. Specifications with fixed effects include day-of-admission fixed effects. CCM calculated from model without fixed effects. Baseline covariates included in all regressions. Sample is all male spells during randomization with complete 18-month follow-up data. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.16 Juvenile Detention Study 3 – Treatment Effect on Number of Re-admissions and Arrests within Given Number of Months, Balanced Panel**

| Months Since Release | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| ITT without Fixed Effects | -0.0421 | -0.0658 | -0.0500 | -0.0761 | -0.1025 | -0.1043 | -0.1256 | -0.1154 | -0.1134 |
| | (0.0299) | (0.0462) | (0.0584) | (0.0695) | (0.0820) | (0.0913) | (0.1019) | (0.1113) | (0.1175) |
| ITT with Fixed Effects | -0.0476 | -0.0869* | -0.0695 | -0.1056 | -0.1421 | -0.1486 | -0.1768 | -0.1729 | -0.1725 |
| | (0.0320) | (0.0493) | (0.0617) | (0.0742) | (0.0875) | (0.0980) | (0.1101) | (0.1201) | (0.1265) |
| CM | 0.644 | 1.228 | 1.746 | 2.210 | 2.660 | 3.107 | 3.536 | 3.950 | 4.343 |
| | | | | | | | | | |
| LATE without Fixed Effects | -0.1655 | -0.2586 | -0.1964 | -0.2991 | -0.4029 | -0.4099 | -0.4937 | -0.4535 | -0.4458 |
| | (0.1171) | (0.1814) | (0.2284) | (0.2719) | (0.3206) | (0.3569) | (0.3987) | (0.4350) | (0.4591) |
| LATE with Fixed Effects | -0.1819 | -0.3323* | -0.2656 | -0.4036 | -0.5434* | -0.5681 | -0.6757* | -0.6610 | -0.6595 |
| | (0.1168) | (0.1803) | (0.2247) | (0.2702) | (0.3191) | (0.3574) | (0.4019) | (0.4382) | (0.4615) |
| CCM | .637 | 1.273 | 1.786 | 2.382 | 2.923 | 3.339 | 3.816 | 4.213 | 4.59 |

Notes: n = 2693. Dependent variable is count of how many times youth returned to JTDC or was re-arrested within X months. Because youth can be sent to the JTDC by a judge within 21 days of the arresting offense, we count any arrest/admission combination as 1 incident if the arrest precedes the admission by 21 days or less. Some individuals have multiple spells in the data, so standard errors are clustered on the individual. ITT is the intention to treat effect, while LATE is the IV estimate for the effect of participation in CBT using random assignment to CBT as an instrument for participation in CBT. LATE operationalizes treatment receipt as spending more than 5% of a stay in a treatment unit. Specifications with fixed effects include day-of-admission fixed effects. CCM calculated from model without fixed effects. Baseline covariates included in all regressions. Sample is all male spells during randomization with complete 18-month follow-up data. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

**Table A.17 Juvenile Detention Study 3 – Treatment Effect on Probability of Re-admission within Given Number of Months, Full Sample**

| Months Since Release | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| ITT without Fixed Effects | -0.0010 | -0.0037 | -0.0016 | -0.0030 | -0.0076 | -0.0216* | -0.0325** | -0.0325** | -0.0338** |
| | (0.0117) | (0.0128) | (0.0128) | (0.0125) | (0.0123) | (0.0127) | (0.0136) | (0.0150) | (0.0170) |
| ITT with Fixed Effects | -0.0054 | -0.0112 | -0.0087 | -0.0102 | -0.0131 | -0.0279** | -0.0385*** | -0.0397** | -0.0430** |
| | (0.0122) | (0.0133) | (0.0132) | (0.0130) | (0.0129) | (0.0133) | (0.0143) | (0.0157) | (0.0181) |
| CM | 0.292 | 0.441 | 0.519 | 0.568 | 0.604 | 0.637 | 0.661 | 0.677 | 0.689 |
| | | | | | | | | | |
| LATE without Fixed Effects | -0.0053 | -0.0183 | -0.0077 | -0.0146 | -0.0363 | -0.1012* | -0.1388** | -0.1379** | -0.1328** |
| | (0.0583) | (0.0638) | (0.0634) | (0.0613) | (0.0587) | (0.0594) | (0.0586) | (0.0641) | (0.0670) |
| LATE with Fixed Effects | -0.0269 | -0.0557 | -0.0430 | -0.0497 | -0.0619 | -0.1287** | -0.1609*** | -0.1636*** | -0.1643** |
| | (0.0578) | (0.0629) | (0.0622) | (0.0604) | (0.0582) | (0.0589) | (0.0577) | (0.0627) | (0.0669) |
| CCM | .24 | .413 | .513 | .574 | .613 | .694 | .739 | .748 | .763 |
| N | 5713 | 5709 | 5696 | 5683 | 5502 | 4983 | 4328 | 3493 | 2693 |

Notes: Dependent variable is indicator for whether youth returned to JTDC within X months. Some individuals have multiple spells in the data, so standard errors are clustered on the individual. Probit results similar to linear probability model shown. ITT is the intention to treat effect, while LATE is the IV estimate for the effect of participation in CBT using random assignment to CBT as an instrument for participation in CBT. LATE operationalizes treatment receipt as spending more than 5% of a stay in a treatment unit. Specifications with fixed effects include day-of-admission fixed effects. CCM calculated from model without fixed effects. Baseline covariates included in all regressions. Sample is all male spells during randomization, but each month's regression uses only spells that are observed for the entire X-month follow-up period. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.18 Juvenile Detention Study 3 – Treatment Effect on Probability of Re-admission within Given Number of Months, Full Sample, First Spell in Randomization Only**

| Months Since Release | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| ITT without Fixed Effects | -0.0123 | -0.0181 | -0.0161 | -0.0147 | -0.0126 | -0.0296* | -0.0399** | -0.0315 | -0.0305 |
| | (0.0154) | (0.0171) | (0.0175) | (0.0174) | (0.0173) | (0.0175) | (0.0180) | (0.0191) | (0.0207) |
| ITT with Fixed Effects | -0.0172 | -0.0259 | -0.0243 | -0.0258 | -0.0203 | -0.0392** | -0.0523*** | -0.0424** | -0.0443** |
| | (0.0169) | (0.0190) | (0.0191) | (0.0189) | (0.0189) | (0.0188) | (0.0194) | (0.0204) | (0.0221) |
| CM | 0.253 | 0.384 | 0.458 | 0.502 | 0.539 | 0.575 | 0.605 | 0.626 | 0.645 |
| | | | | | | | | | |
| LATE without Fixed Effects | -0.0583 | -0.0855 | -0.0759 | -0.0685 | -0.0582 | -0.1339* | -0.1659** | -0.1284 | -0.1213 |
| | (0.0727) | (0.0805) | (0.0819) | (0.0808) | (0.0795) | (0.0791) | (0.0756) | (0.0781) | (0.0822) |
| LATE with Fixed Effects | -0.0765 | -0.1148 | -0.1073 | -0.1120 | -0.0884 | -0.1676** | -0.2050*** | -0.1611** | -0.1695** |
| | (0.0687) | (0.0770) | (0.0771) | (0.0751) | (0.0751) | (0.0743) | (0.0713) | (0.0722) | (0.0794) |
| CCM | .245 | .419 | .523 | .58 | .616 | .708 | .74 | .732 | .731 |
| N | 2984 | 2983 | 2978 | 2971 | 2924 | 2779 | 2569 | 2229 | 1860 |

Notes: Dependent variable is indicator for whether youth returned to JTDC within X months. Probit results similar to linear probability model shown. ITT is the intention to treat effect, while LATE is the IV estimate for the effect of participation in CBT using random assignment to CBT as an instrument for participation in CBT. LATE operationalizes treatment receipt as spending more than 5% of a stay in a treatment unit. Specifications with fixed effects include day-of-admission fixed effects. CCM calculated from model without fixed effects. Baseline covariates included in all regressions. Sample is the first spell during the randomization period for each male youth. Each month's regression uses only spells that are observed for the entire X-month follow-up period. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.19 Juvenile Detention Study 3 – Treatment Effect on Number of Re-admissions in Given Number of Months, Full Sample**

| Months Since Release | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| ITT without Fixed Effects | -0.0037 | -0.0041 | 0.0009 | -0.0105 | -0.0226 | -0.0357 | -0.0742* | -0.1152** | -0.1323** |
| | (0.0140) | (0.0201) | (0.0247) | (0.0284) | (0.0323) | (0.0379) | (0.0445) | (0.0544) | (0.0656) |
| ITT with Fixed Effects | -0.0110 | -0.0208 | -0.0201 | -0.0332 | -0.0482 | -0.0644 | -0.1149** | -0.1573*** | -0.1843*** |
| | (0.0147) | (0.0210) | (0.0257) | (0.0296) | (0.0338) | (0.0393) | (0.0467) | (0.0567) | (0.0688) |
| CM | 0.334 | 0.607 | 0.823 | 1.016 | 1.190 | 1.372 | 1.533 | 1.678 | 1.789 |
| | | | | | | | | | |
| LATE without Fixed Effects | -0.0186 | -0.0203 | 0.0043 | -0.0517 | -0.1079 | -0.1673 | -0.3173* | -0.4894** | -0.5201** |
| | (0.0700) | (0.1004) | (0.1230) | (0.1394) | (0.1540) | (0.1765) | (0.1900) | (0.2315) | (0.2574) |
| LATE with Fixed Effects | -0.0544 | -0.1032 | -0.0992 | -0.1617 | -0.2281 | -0.2975* | -0.4798** | -0.6489*** | -0.7045*** |
| | (0.0697) | (0.0993) | (0.1207) | (0.1374) | (0.1526) | (0.1735) | (0.1876) | (0.2251) | (0.2522) |
| CCM | .271 | .557 | .755 | 1.022 | 1.229 | 1.413 | 1.662 | 1.875 | 2.011 |
| N | 5713 | 5709 | 5696 | 5683 | 5502 | 4983 | 4328 | 3493 | 2693 |

Notes: Dependent variable is count of how many times youth returned to JTDC within X months. Some individuals have multiple spells in the data, so standard errors are clustered on the individual. ITT is the intention to treat effect, while LATE is the IV estimate for the effect of participation in CBT using random assignment to CBT as an instrument for participation in CBT. LATE operationalizes treatment receipt as spending more than 5% of a stay in a treatment unit. Specifications with fixed effects include day-of-admission fixed effects. CCM calculated from model without fixed effects. Baseline covariates included in all regressions. Sample is all male spells during randomization, but each month's regression uses only spells that are observed for the entire X-month follow-up period. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.20 Juvenile Detention Study 3 – Treatment Effect on Number of Re-admissions and Arrests within Given Number of Months, Full Sample**

| Months Since Release | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| ITT without Fixed Effects | -0.0020 | -0.0109 | 0.0065 | -0.0118 | -0.0360 | -0.0533 | -0.1424* | -0.1652* | -0.1134 |
| | (0.0203) | (0.0315) | (0.0412) | (0.0483) | (0.0560) | (0.0657) | (0.0790) | (0.0975) | (0.1175) |
| ITT with Fixed Effects | -0.0073 | -0.0262 | -0.0190 | -0.0422 | -0.0619 | -0.0845 | -0.1887** | -0.2317** | -0.1725 |
| | (0.0215) | (0.0326) | (0.0421) | (0.0495) | (0.0582) | (0.0684) | (0.0826) | (0.1026) | (0.1265) |
| CM | 0.5920 | 1.146 | 1.641 | 2.116 | 2.583 | 3.075 | 3.545 | 3.988 | 4.343 |
| | | | | | | | | | |
| LATE without Fixed Effects | -0.0100 | -0.0547 | 0.0324 | -0.0583 | -0.1720 | -0.2495 | -0.6094* | -0.7017* | -0.4458 |
| | (0.1013) | (0.1570) | (0.2050) | (0.2374) | (0.2666) | (0.3059) | (0.3377) | (0.4130) | (0.4591) |
| LATE with Fixed Effects | -0.0364 | -0.1296 | -0.0935 | -0.2056 | -0.2927 | -0.3899 | -0.7881** | -0.9560** | -0.6595 |
| | (0.1016) | (0.1537) | (0.1979) | (0.2299) | (0.2624) | (0.3022) | (0.3327) | (0.4065) | (0.4615) |
| CCM | .555 | 1.136 | 1.601 | 2.187 | 2.756 | 3.235 | 3.836 | 4.265 | 4.59 |
| N | 5711 | 5707 | 5694 | 5681 | 5500 | 4981 | 4327 | 3492 | 2692 |

Notes: Dependent variable is count of how many times youth returned to JTDC or was re-arrested within X months. Because youth can be sent to the JTDC by a judge within 21 days of the arresting offense, we count any arrest/admission combination as 1 incident if the arrest precedes the admission by 21 days or less. Some individuals have multiple spells in the data, so standard errors are clustered on the individual. ITT is the intention to treat effect, while LATE is the IV estimate for the effect of participation in CBT using random assignment to CBT as an instrument for participation in CBT. LATE operationalizes treatment receipt as spending more than 5% of a stay in a treatment unit. Specifications with fixed effects include day-of-admission fixed effects. CCM calculated from model without fixed effects. Baseline covariates included in all regressions. Sample is all male spells during randomization, but each month's regression uses only spells that are observed for the entire X-month follow-up period. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.21 Treatment Effect on Total Criminal Incidents Per Youth Per Year across All 3 Studies**

| | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean |
|---|---|---|---|---|
| | | **Panel A. Pooled** | | |
| Treatment | 1.5460 | -0.0905** | -0.2398** | 1.0090 |
| | | (0.0397) | (0.1048) | |
| | | **Panel B. Pooled with Study Interactions** | | |
| Treatment | 0.6150 | -0.1019** | -0.2037** | 0.6890 |
| | | (0.0488) | (0.0969) | |
| Treatment*Study 1 | 0.6990 | 0.0310 | 0.0321 | 0.6570 |
| | | (0.0671) | (0.1467) | |
| Treatment*Study 3 | 3.0950 | 0.0003 | -0.1965 | 3.3850 |
| | | (0.1034) | (0.3691) | |
| P, test all interactions = 0 | | 0.8872 | 0.8266 | |
| | | **Panel C. Pooled with Study Interactions, SD units** | | |
| Treatment | 0.0000 | -0.0750** | -0.1499** | 0.0530 |
| | | (0.0361) | (0.0717) | |
| Treatment*Study 1 | 0.0000 | 0.0278 | 0.0356 | -0.0320 |
| | | (0.0480) | (0.1041) | |
| Treatment*Study 3 | 0.0000 | 0.0362 | -0.0036 | 0.1090 |
| | | (0.0516) | (0.1610) | |
| P, test all interactions = 0 | | 0.7581 | 0.9354 | |

Notes: n = 7496. Sample is all observations for studies 1 and 2, plus the observations in study 3 with at least 18 months of follow-up data and non-missing arrest data. Outcome is number of total criminal incidents in one year. For study 1, this is total arrests in year 1. For study 2, it is total arrests during the program period scaled to a one-year period (total in 19 months / 19 * 12). For study 3, it is total number of readmissions and rearrests in the 12 months post-release. Panel A pools all studies together. Panel B uses Study 2 as the reference group in a pooled regression and includes study-by-treatment interactions for studies 1 and 3 (CM and CCM columns show control or control complier mean for the study in that row). Panel C replicates Panel B but using a version of the dependent variable that is standardized for each study using that study's control group distribution. Baseline covariates and randomization block fixed effects included in all models (see text). Since some youth are in multiple studies, standard errors in parentheses are clustered on individual. * p<0.1, ** p<0.05, *** p<0.01.

**Table A.22 Test of Candidate Mediating Mechanisms, Study 1 Sample (BAM 2009-10 Cohort)**

| | | School Engagement (2009-10) | | School Engagement (2010-11) | | Graduation by June 2015 | |
|---|---|---|---|---|---|---|---|
| Candidate mediating measure (Z-score form, normalized to control group distribution) | Effect of BAM participation on candidate mediator | Association of mediator with school engagment among controls | % BAM effect on school engagement explained by this mechanism | Association of mediator with school engagment among controls | % BAM effect on school engagement explained by this mechanism | Association of mediator with graduation among controls | % BAM effect on graduation explained by this mechanism |
| Social Capital /Mentoring | 0.0173 | -0.0353 | -0.44% | -0.0203 | -0.40% | -0.002 | 0.00% |
| (Participation n = 999) (Outcomes n = 428) | (0.1243) | (0.0216) | (-11.97%, 10.1%) | (0.0216) | (-6.27%, 4.09%) | (0.0195) | (-32.52%, 18.76%) |
| Perceived Returns to Schooling | -0.0219 | 0.0154 | -0.22% | 0.0420* | -0.90% | -0.0077 | 0.28% |
| (Participation n = 794) (Outcomes n = 340) | (0.1527) | (0.0224) | (-14.09%, 9.53%) | (0.0225) | (-14.17%, 9.09%) | (0.0241) | (-39.26%, 32.23%) |
| Social skills | 0.1316 | 0.0014 | 0.15% | -0.0019 | -0.30% | -0.0036 | -0.70% |
| (Participation n = 1081) (Outcomes n = 446) | (0.1196) | (0.0169) | (-5.97%, 7.64%) | (0.0187) | (-5.54%, 4.44%) | (0.0198) | (-43.35%, 33.3%) |
| Grit | 0.1145 | 0.0689*** | 5.78% | 0.0471** | 5.40% | -0.0157 | -2.52% |
| (Participation n = 975) (Outcomes n = 417) | (0.1308) | (0.0220) | (-11.31%, 33.56%) | (0.0215) | (-5.18%, 15.03%) | (0.0218) | (-61.46%, 29.77%) |

Notes: The following are the specific questions for each category. 1. Social Capital/Mentoring: have at least one teacher or adult in school I can talk to if I have a problem. 2. Schooling: classes are useful preparation for future; high school teaches valuable skills; working hard in school matters for work force; what we learn in class is useful for future. 3. Social skills: I can always find a way to help end arguments; I listen carefully to what other people say about me; I'm very good at working with other students; I'm good at helping people. 4. Grit: I finish whatever I begin; I am a hard worker.

First column of results presents coefficient from IV analysis of BAM participation effect on the candidate mediating mechanism measure listed in the row label at left, which comes from a survey of youth in CPS carried out by the Chicago Consortium of School research in 2011 (see text). Second column presents the results of a non-experimental regression of the candidate mediator against the school engagement index (outcome), using just data from the control group. Both sets of models control for the same baseline covariates and randomization block fixed effects as in the main analyses with heteroskedasticity-robust standard errors in parentheses. Third column multiplies point estimate from column 1 by point estimate in column 2 and then divides by the estimated IV effect of BAM participation on that outcome taken from Table 4. Confidence intervals are bootstrapped using 1,999 replications. Remaining columns of the table are constructed analogously. * p<0.10, ** p<0.05, *** p<0.01.

**Table A.23 Test of Candidate Mediating Mechanisms, Study 1 Sample (BAM 2009-10 Cohort)**

| | | Violent Crime Arrests (2009-10) | | Total Arrests (2009-10) | |
|---|---|---|---|---|---|
| Candidate mediating measure (Z-score form, normalized to control group distribution) | Effect of BAM participation on candidate mediator | Association of mediator with violent crimes among controls | % BAM effect on violent crime arrests explained by this mechanism | Association of mediator with total arrests among controls | % BAM effect on total arrests explained by this mechanism |
| Social Capital /Mentoring<br>(Participation n = 999) (Outcomes n = 428) | 0.0173<br>(0.1243) | -0.0139<br>(0.0132) | 0.24%<br>(-8.92%, 9.87%) | -0.0382<br>(0.0381) | 0.37%<br>(-13.94%, 14.57%) |
| Perceived Returns to Schooling<br>(Participation n = 794) (Outcomes n = 340) | -0.0219<br>(0.1527) | -0.0235<br>(0.0210) | -0.60%<br>(-15.73%, 15.8%) | 0.0503<br>(0.0658) | 0.59%<br>(-28.71%, 29.19%) |
| Social skills<br>(Participation n = 1081) (Outcomes n = 446) | 0.1316<br>(0.1196) | 0.0023<br>(0.0148) | -0.36%<br>(-12.99%, 10.12%) | 0.0433<br>(0.0449) | -3.05%<br>(-34.44%, 13.16%) |
| Grit<br>(Participation n = 975) (Outcomes n = 417) | 0.1145<br>(0.1308) | 0.0001<br>(0.0172) | 0.00%<br>(-12.93%, 11.92%) | 0.0543<br>(0.0544) | -3.32%<br>(-38.91%, 22.76%) |

Notes: The following are the specific questions for each category. 1. Social Capital/Mentoring: have at least one teacher or adult in school I can talk to if I have a problem. 2. Schooling: classes are useful preparation for future; high school teaches valuable skills; working hard in school matters for work force; what we learn in class is useful for future. 3. Social skills: I can always find a way to help end arguments; I listen carefully to what other people say about me; I'm very good at working with other students; I'm good at helping people. 4. Grit: I finish whatever I begin; I am a hard worker.

First column of results presents coefficient from IV analysis of BAM participation effect on the candidate mediating mechanism measure listed in the row label at left, which comes from a survey of youth in CPS carried out by the Chicago Consortium of School Research in 2011 (see text). Second column presents the results of a non-experimental regression of the candidate mediator against the school engagement index (outcome), using just data from the control group. Both sets of models control for the same baseline covariates and randomization block fixed effects as in the main analyses with heteroskedasticity-robust standard errors in parentheses. Third column multiplies point estimate from column 1 by point estimate in column 2 and then divides by the estimated IV effect of BAM participation on that outcome taken from Table 4. Confidence intervals are bootstrapped using 1,999 replications. Remaining columns of the table are constructed analogously. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

**Table A.24 Design of Decision-Making Experiment Carried Out With Sub-Sample of Study (BAM 2013-14 Cohort)**

| | Randomized to BAM | | Randomized to Control | |
|---|---|---|---|---|
| | Slow down? | Reflect? | Slow down? | Reflect? |
| Condition 1<br>No delay | Yes | Yes | | |
| Condition 2<br>Delay (distraction) - "partial CBT" manipulation | Yes | Yes | Yes | |
| Condition 3<br>Delay plus reflection - "CBT" manipulation | Yes | Yes | Yes | Yes |
| Condition 4<br>Delay plus rumination -"anti-CBT" manipulation | Yes | | Yes | |

**Table A.25 Effect of Treatment on Decision-Making and Automaticity, Study 2 (BAM 2013-14 Cohort)**

| | Time (seconds) (n=302) | | Time (no outliers) (n=295) | | Log time (n=302) | |
|---|---|---|---|---|---|---|
| | Control Complier Mean | Effect of Participation (IV) | Control Complier Mean | Effect of Participation (IV) | Control Complier Mean | Effect of Participation (IV) |
| All Conditions Pooled | 3.742 | 1.1269* | 3.207 | 1.2691*** | 0.969 | 0.3264** |
| (n = 302) | | (.6661) | | (.4648) | | (.1338) |
| Condition 1 | 4.787 | 2.2781 | 3.889 | 2.2129*** | 1.102 | 0.5955** |
| No delay (n = 117) | | (1.5609) | | (.8495) | | (.2608) |
| Condition 2 | 3.036 | 0.1947 | 2.551 | 0.6802 | 0.860 | 0.1076 |
| Delay (n = 60) | | (.9237) | | (.6226) | | (.2239) |
| Condition 3 | 3.795 | 0.7649 | 3.403 | 1.1575 | 0.999 | 0.2063 |
| Delay plus reflection (n = 63) | | (1.0675) | | (1.0033) | | (.2447) |
| Condition 4 | 2.003 | 0.9083 | 2.140 | 0.7721 | 0.669 | 0.3121 |
| Delay plus rumination (n = 62) | | (.8635) | | (.802) | | (.2335) |
| Conditions 1-3 Pooled | 4.340 | 1.0733 | 3.618 | 1.3079** | 1.082 | 0.3034* |
| (n = 240) | | (.8746) | | (.5817) | | (.1686) |

Notes: Table presents results from administering iterated dictator game to sub-sample of youth in BAM study 2. Outliers defined as taking over 20 seconds to make decision on take amount. Regression specification includes baseline covariates and randomization block fixed effects as in main analyses. Heteroskedasticity-robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

**Table A.26 Becoming a Man Study 1 – Estimated Benefits Per Participant**

| | Low-end Estimate | High-end Estimate |
|---|---|---|
| | From Year 1 Crime Reduction | |
| Savings to Potential Victims | 4,615 | 32,918 |
| | (3,161) | (21,381) |
| Savings to Government | 720* | 1,268* |
| | (407) | (718) |
| Subtotal | 5,335* | 34,186 |
| | (3,238) | (21,473) |
| | From Increased High School Graduation | |
| Earnings Increase to Participant | 1,011 | 5,617** |
| | (633) | (2,644) |
| Cost of Additional Schooling | -294 | -678** |
| | (202) | (323) |
| Subtotal | 716 | 4,939** |
| | (455) | (2,338) |
| | Total | |
| | 6,051* | 39,125* |
| | (3,282) | (21,701) |
| Costs per participant | $1,100 | $1,100 |
| Benefits/Costs | 6/1 | 36/1 |

Notes: Table assigns a social cost to each crime (top panel) and a social benefit to each high school graduate (bottom panel), then estimates an individual level program benefit with an IV regression using social cost as the dependent variable. All estimates reported in 2010 dollars. The low estimate column uses lower-bound participation rates, the cost of the cost of crimes to victims from Miller, Cohen and Wiersema (1996) with the cost of homicide trimmed by half, the measure of graduating from CPS only, and the lower range of estimated earnings and health benefits from an additional year of education in the literature. The high estimate column uses participation as reported, the costs of crime from the contingent valuation surveys in Cohen, et al. (2004), the measure of graduation that assumes all verified out-of-district transfers graduate, and the higher end of estimated earnings and health benefits from a year of education. Both columns assign each graduate the cost of one extra year of instruction in CPS and each offender the cost of each arrest to the criminal justice system. See Appendix C Section V for details. Baseline covariates and randomization block fixed effects included. Heteroskedasticity-robust standard errors in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.