# TECHNICAL APPENDIX

## I. OVERVIEW

This appendix presents additional technical details about our team's series of randomized controlled trials (RCTs) of the Chicago-area non-profit Youth Guidance's Becoming a Man (BAM) program, and what we have learned about the program's scope and scale.

BAM is an in-school program designed to help youth develop non-academic skills and, more generally, to encourage youth to reflect on their decision-making heuristics, or to promote both meta-cognition (to "think about thinking") and deeper emotional connections. Between 2009 and 2015 our team carried out a total of 4 RCTs with 9,804 middle- and high-school aged youth in the Chicago Public Schools (CPS) system.

As discussed in the main text, results for the initial RCTs of BAM showed very large behavioral impacts, suggesting the program can solve the challenge of scope (that is, it changes a sufficiently large set of important decisions to lead to large changes in important policy outcomes like high school dropout and violence involvement). But as the program has expanded over time and served more youth, the estimated impact of BAM on our key outcomes – total arrests, violent-crime arrests and schooling – seem to have become more variable. It should be noted that our statistical power is somewhat limited; that is, in the later RCTs carried out when the program is serving a relatively larger number of youth citywide in Chicago, we cannot reject the null hypothesis that impacts are zero, but we also cannot reject the null hypothesis that results are the same across all of the different experiments. This raises the possibility that while BAM has solved the scope challenge, it may have more difficulty in solving the problem of scale.

In what follows we describe each of the RCTs in more detail, discuss the findings from each individual RCT and how they compare to one another, and present our results from trying to understand the explanation for why results seem to differ across experiments. Specifically, we explore the extent to which variability in the population served, implementation fidelity, school settings, and neighborhood contexts – as well as the interaction of these potential moderators of treatment effects – do or do not explain differential treatment effects observed across studies.

## II. BECOMING A MAN

Becoming a Man is a group mentoring program developed by Youth Guidance that tries to help youth strengthen their meta-cognition – that is, their thinking about their own thinking. BAM is related to what clinical psychologists call cognitive behavioral therapy (CBT). Participants have the opportunity to participate in one-hour, once-per-week group sessions held during the school day. The intervention is delivered in groups, to help hold down costs, but the groups are intended to be kept small to help develop relationships (assigned groups are not more than fifteen students, with a realized average of eight). Students are excused from a class to participate, which is a draw for some youth.

The program was developed by Youth Guidance about a decade before the first RCT of the program that our team carried out in academic year (AY) 2009–10. The program was operating in a single Chicago high school and a few elementary schools before being taken to scale for the first RCT in AY2009-10. The BAM curriculum itself has substantially changed since this first RCT. In 2009-10, the curriculum was designed to be one year long, with about 30 sessions. Since 2013, and in our most recent RCTs, the curriculum is over two years (up to 45 sessions).

In addition, between our 2009-10 study of BAM and our 2013-15 studies of BAM, Youth Guidance made substantial efforts to provide additional training and supervision of counselors to best accommodate implementation with fidelity to the model at large-scale in Chicago. Namely, delivery of BAM in AY2013-14 included for the first time the hiring of BAM supervisors (with a staff-to-supervisor ratio of five-to-one), the development of infrastructure support and capacity building roles, a fidelity monitoring dashboard, and additional efforts to develop and manualize the curriculum. In turn, the costs of the program increased from about $1,100 per participant in 2009 to about $1,850 per participant per year in the most recent RCTs (or about $3,700 per participant for the entire two-year program period).

Table 1 illustrates a few of the key types of activities included in the BAM curriculum and provides a brief description of each selected activity. The program has a program-specific manual and facilitator's guide and is delivered by college-educated men.[1] Youth Guidance prioritizes hiring counselors who are able to keep youth engaged and aims to hire people from neighborhoods similar to those in which they would be working.

The curriculum includes standard elements of CBT, such as a common structure to most sessions that starts with a "check-in." Youth sit in a circle with the counselor, who reflects on how things in his life are going in various domains. The youth then follow suit. This activity is an example of "retrospective / introspective" activities [2], which include various efforts to get youth to talk about the things they are doing well and areas in which they still need to still improve. Youth discuss both their perception of self and their perceptions of peers on these two dimensions.

Another type of activity in the BAM curriculum is "immersive or experiential," of which one example is called the stick. Youth are divided into two groups and lined up facing each other. They are told to put their arms out chest high and extend their index fingers, and the counselor then lays a 10- or 15-foot plastic pipe across everyone's fingers. The group is told that they must lower the pipe to the floor but their fingers must be touching the pipe at all times. This leads everyone to put upward pressure on the pipe, which makes it go up rather than down. As youth become immersed in the activity, they can lose themselves in the moment and become frustrated, blaming each other rather than recognizing that each of them contributes to the problem—and that they could help solve the problem themselves by trying to coordinate and lead the group.

Other types of activities included in the BAM curriculum are "role-playing" and "stories and discussions." For example, in the $10 role-play activity, students act out a scene in which one of them has borrowed money from another but then never paid it back. The youth act out how they would respond and then the group discusses what happened and why, and what might have led to

[1] While not required, Youth Guidance has a preference for training in psychology or social work when selecting program providers.

[2] Activity labels were coded by the research team, not by Youth Guidance, for research purposes.

a better outcome. Stories include the elephant and the rhino, in which two large animals are very persistent in their refusal to make way for the other, to both their detriments.

The program also does some “skill-building.” This includes lessons in muscle relaxation, deep breathing, and channeling anger productively. It also includes cognitive thought replacement, a CBT element that asks youth to identify and replace problematic or false beliefs. Finally, the curriculum includes a discussion of different conceptions of masculinity and some general values like the importance of integrity and personal accountability. It also takes youth on field trips to local colleges to highlight the value of education, and, by putting youth in regular contact with a pro-social adult, has a mentoring component as well.

In addition to the CBT components of the curriculum, BAM counselors also aim to develop additional social-emotional skills through the program including impulse control/emotional self-regulation, social information processing, future orientation, and integrity. Weekly group sessions are structured around six core values to develop these social-emotional skills: integrity, accountability, self-determination, positive anger expression, visionary goal-setting, and respect for womanhood. These values are delivered via a multi-faceted approach using various youth engagement activities. By focusing sessions on CBT and these core values, Youth Guidance hopes to ultimately improve non-cognitive skill development and academic engagement, with the immediate goals of reducing school suspensions and increasing school attendance, the intermediate goals of improving academic achievement and decreasing arrests, and the long-term goals of improving graduation rates and reducing involvement in violence among participants.

BAM is, in short, a program with multiple elements and hence multiple potential mechanisms of action through which it may change youth behavior and outcomes. We return below to the question of what can be surmised about what the most important mechanisms might be.

## III. RESEARCH DESIGN AND STUDY SAMPLES

### A. Overview of all studies

Table 2 presents a summary table of all studies to date on the Becoming a Man program: *BAM 1* refers to the original RCT conducted of a one-year BAM curriculum in academic year (AY) 2009-10. *BAM 2* refers to a second study of a two-year curriculum conducted from AY2013-15 in which students were assigned to BAM or a control group in 2013. *BAM 2x2* refers to a third study of the two-year curriculum conducted in AY2013-15 that was part of a larger 2x2 factorial experiment that aimed to measure the synergistic effects of academic and non-academic supports. The academic support is the high-intensity tutoring, delivered by SAGA Innovations, which was reported on by Cook, et al. (2015); results for those interaction effects will be reported on in a separate paper. *BAM Expansion* refers to groups of students who were randomized in 2014, or the second year of the 2013-15 study, but for whom we do not have the second year of participation data from AY2015-16.

We analyze various subsamples depending on the set of schools the students attended and the year in which students were randomized, as depicted in Table 2. The shaded cells indicate when each study was in operation, delivering the program to youth, with the sample size (N) reported in the relevant cells.  Each of these samples are described in greater detail below.

We note that all control group students in our studies were allowed to participate in status quo school and community services. Neither Youth Guidance nor the evaluation team thoroughly documented all of the additional services available at the school. However, based on our knowledge of the schools, we do not currently know of similar programs that were being administered at any of the schools during our study years.

## B. Overview of individual studies

### 1. BAM 1: Randomization in AY2009-10

In AY2009-10, the University of Chicago Crime Lab conducted a study of the BAM program in 18 CPS schools in Chicago's distressed south and west sides. In this RCT, 2,740 male youth in 7th through 10th grade were assigned to one of two groups for one academic year—to BAM, or to a control group that received status quo school and community services.

It is important to note that the BAM intervention in our 2009 study differed slightly from the intervention studied in our 2013-15 evaluations. The most significant difference is that the curriculum for BAM 1 only spanned one year, while the curricula for the remaining BAM evaluations spanned two years.[3] Other differences include modifications to the curriculum itself and different training provided to counselors.

Table 3 shows the sample of youth in BAM 1 represent a population that has high levels of criminal justice involvement and disengagement from school. Youth were between 15 and 16 years old at baseline. About 70% were African-American, and 30% were Hispanic. Youth had an average GPA of 1.7 on a 4-point scale, and about a third of youth had an arrest at baseline.

### 2. BAM 2: Randomization in AY2013-14

In AY2013-14, the University of Chicago Crime Lab began its second study of BAM by randomizing 2,064 male 9th and 10th graders to one of two groups in nine CPS high schools for two academic years—to BAM, or to a control group that received status quo school and community services. We note that one of these nine schools did not continue the BAM curriculum in AY2014-15, and we thus drop this school when looking at impacts during year 2 of the study.[4]

Table 3 indicates students in the BAM 2 study were slightly younger than in BAM 1, and had lower levels of criminal justice involvement and poorer school performance. In this study, youth were about 15 years old at baseline. Reflecting the composition of their neighborhoods, around two-thirds of youth are black and the remainder Hispanic. Youth had an average GPA of 2.1 on a 4-point scale, and about a quarter of youth had an arrest at baseline.

### 3. BAM 2x2: Randomization in AY2013-14

[3] For more information on the 2009 BAM study, please see Heller, et al. (2017).

[4] One of the schools in our BAM 2 study decided that they did not want the program to continue in AY2014-15. We include this school when looking at impacts for year 1, but drop this school when looking at impacts for year 2.

The “BAM 2x2” study was part of a larger NIH-funded effort to understand the degree to which developing academic and non-academic skills has synergistic effects on youth outcomes. Randomization occurred in twelve CPS high schools among 2,633 male 9th and 10th grade students. The study sample was randomized to one of four groups—to BAM, a high-intensity math tutoring program, both BAM and the high-intensity math tutoring program, or a control group that was offered neither program but received status quo school and community services. In order to be able to populate all four “cells” of the randomization, we wound up having to work in a slightly larger set of high schools compared to other BAM studies. We discuss the role of this shift in school sample in explaining differences in impact estimates in more detail below.

In the results we present here, we compare all students who did not receive BAM (all students who were randomized to receive math tutoring or to the control group) against all students who did receive BAM (all students who were randomized to receive both programs, or to just receive BAM).[5] Youth randomized to treatment were to receive BAM for two academic years. We drop one school from our BAM 2x2 analyses due to a “broken experiment” in that school; that is, the program providers in the school failed to follow the random assignment and (we fear) incompletely documented which youth actually received the program.[6]

Table 3 describes the population of youth for our BAM 2x2 sample. Over 45% of youth in our sample are African-American, and over 45% are Hispanic. This differs substantially from our BAM 1 and BAM 2 samples, where about two-thirds of study participants are African-American, and one-third are Hispanic. About 86% are eligible for free lunch. The average GPA for these youth was 2.1 on a 4-point scale, and about 20% of youth had an arrest at baseline. Fewer youth had an arrest at baseline in this sample compared to our BAM 1 and BAM 2 samples.

### 4. BAM Expansion: Randomization in AY2014-15

During the AY2014-15, we randomized an additional cohort, mostly of incoming 9th graders, to receive BAM in both BAM 2x2 and BAM 2 schools. As our study ended in AY2014-15, we do not have data for these youth from their second year of intervention, which would have occurred in AY2015-16. Consequently, we report data for these youth for one year of intervention to date. We return to this issue when pooling the estimates from all of our BAM samples.

[5] There may be some concerns with collapsing all students who receive BAM in our analyses, as we may be at risk for underreporting BAM effects across individual study results if the synergistic effects of BAM and the math tutoring program somehow diminish BAM effects. However, based on our preliminary analysis of the full two-by-two experiment, we do not believe this to be true.

[6] We received notification from Youth Guidance leadership post-study that a BAM counselor in one school allowed control students in BAM sessions without documentation in his attendance data. When conducting analysis for this study in fall 2016, we received names from this counselor of additional control students he served in both AY2013-14 and AY2014-15. In addition, there were two other counselors at this school for whom we do not know whether they served control students or not. In our own internal program data, we calculate that this school served four BAM control students out of a total population of 90 students served over both program years. Using the counselor’s additional rosters, we find that 36 BAM control students were served out of a total sample of 130 students served over both program years. We do not have similar documentation of control students served by the other two counselors. Due to the failure of randomization at this school we drop it from our BAM 2x2 analyses.

Table 3 describes the population of youth for our BAM expansion study sample. Of the 2,367 youth who are in this sample, about 60% of youth are African-American and about 35% are Hispanic. The average age of students is 14.5, and about 88% are eligible for free or reduced lunch. The average GPA for these youth is 2.3 on a 4-point scale, and about 20% had an arrest at baseline. We again note that fewer youth had an arrest at baseline in this sample compared to our BAM 1 and BAM 2 samples, and that students are younger in this sample.

## IV. DATA

Our study relies on several data sources: (1) longitudinal student-level records from administrative data collected by the Chicago Public Schools (CPS), Chicago Police Department (CPD), and Illinois State Police (ISP); and (3) program provider data from Youth Guidance. Our team also conducted in-person observations in a random sample of BAM sessions to monitor and document program implementation.

### A. Administrative Data

#### 1. Chicago Public Schools (CPS) Data

Our first source of administrative government data come from longitudinal student-level records maintained by CPS. Participants are matched to school administrative records using their unique CPS student ID using Stata. These CPS student records include whether the student has a learning disability; month and year of birth; race/ethnicity; eligibility for free and reduced price lunch; course grades in each subject; enrollment status; absences; and disciplinary actions and suspensions. For all of our BAM studies, we create a summary index of school engagement to both reduce the risk of false positives, and to improve statistical power to detect effects for outcomes within a "family" of outcomes that are expected to move in a similar direction. This index is an average of three Z-scored variables: GPA at the end of the school year, days present during a school year, and enrollment status at the end of the year (i.e. whether a student was enrolled at the end of the school year).[7]

Student participation files are initially linked to CPS administrative data in Stata using a unique CPS student ID. Demographic data, such as birth date, learning disability status, and race, are then added from a CPS enrollment file. We then use this demographic information to connect our CPS data to crime data, as described below.

#### 2. Arrest Records

Our second source of administrative government data for these studies are from Chicago Police Department (CPD) and Illinois State Police (ISP) arrest records.

The CPD data that we utilize include information on the identity of the offender, date and location of the crime event, and the criminal charges (for juvenile as well as adult offenders).

[7] We do not include standardized test scores in our school engagement index because the Chicago Public Schools by design did not administer standardized tests to all grades during the years of our study. In addition, our index does not include administrative records on school disciplinary actions, as we are not certain of the validity of this data.

When recording arrests, CPD uses fingerprint technology to identify individuals. The arrest data should therefore include every CPD arrest of an individual, even if he or she submits an alias at the time of arrest. The data also includes arrests that do and do not result in a conviction. We link CPD data to our study samples using probabilistic matching on first name, last name, gender and date of birth.[8] We use CPD data in all of our BAM studies, with the exception of BAM 1.

In our BAM 1 study we used electronic arrest records ("rap sheets") from the Illinois State Police (ISP), obtained through the Illinois Criminal Justice Information Authority (ICJIA). The ISP records capture arrests in the state going back to 1990 and include arrests of people below the age of majority within the criminal justice system (juvenile arrests), as well as to those who are above the age of majority. Local police departments are required by law to report all juvenile felony arrests to the ISP, and optionally class A and B misdemeanors. ICJIA uses probabilistic matching on name, gender and date of birth to match our study sample to ISP arrest records.[9]

Because previous studies often find more pronounced impacts of policy interventions on violent crimes (particularly impulsive crimes such as assault) than on other crimes (Deming 2011, Evans and Owens 2007, Kling, Ludwig and Katz 2005, Lochner and Moretti 2004, Weiner, Lutz and Ludwig 2009) and because associated social harms are so varied across crime types, we examine arrests separately for different offense categories. For each arrest incident, we select the most severe charge associated with the incident. In most cases this is a charge recorded at the time of arrest, although occasionally the State's Attorney files a charge more severe than those originally recorded at the police station. We classify crimes as violent, property, drug, and other:
*Violent crimes* include murder, rape, assault, robbery, threats/harassment, and kidnapping.
*Property crimes* include larceny, burglary, and auto theft.
*Drug crimes* include possession or dealing charges.
*Other crimes* include trespassing, fencing, bribery, animal cruelty, weapons violations, DUIs, disobeying or avoiding law enforcement, disorderly conduct, arson, prostitution, criminal neglect, parole violations, underage or public drinking, vandalism, and miscellaneous offenses.

We exclude from our analysis motor vehicle offenses, including driving with a suspended license, reckless driving, and other driving/traffic related offenses.

### B. Program Provider Data

Another source of data comes from Youth Guidance. Data we receive directly from Youth Guidance includes program attendance data, which provides us with program attendance data for each BAM session held. For our 2013-15 evaluations, BAM counselors keep a log of each group session they lead, which documents the lesson and core values that were taught during that

[8] In our studies, potential matches to CPD data were reviewed and classified using a machine learning algorithm as well as an additional manual review of borderline cases. The resulting arrest-level data was categorized according to the offense type using a combination of the FBI's Uniform Discipline Code and the accompanying statute description from the Chicago Police Department. The data were then aggregated to the student-level and merged onto the analytic file using the CPS student ID variable.

[9] Once the research team received ICJIA arrest data for the BAM 1 study, the data was then categorized according to the offense type using a combination of FBI's Uniform Discipline Code and the accompanying statute description from the Chicago Police Department. The data was then aggregated to the student-level and merged onto the analytic file using a unique CPS student ID.

session, along with a list of students that attended. In addition, BAM counselors documented each individual counseling session, as well as any field trips or out-of-school activities each student was engaged in. Counselors collected this attendance data via a tablet computer and an internal electronic client information system developed by Youth Guidance. Counselors were mandated to fill out this information.

In addition, for these evaluations we receive measures of fidelity of implementation for each counselor based on assessments by Youth Guidance leadership, specifically of counselor and school quality ratings for each school.

### C. In-person Observations

To monitor program implementation during the study period we conducted in-person observations of a random subset of over 90 BAM sessions during AY2014-15. We selected a random subset of sessions, and had a group of graduate and undergraduate research assistants conduct systemic observations of BAM sessions using a standardized observation rubric designed to record information about implementation completeness, quality, and fidelity to the prescribed manual. The BAM rubrics examined constructs related to transitions into and out of BAM sessions, student engagement, activities that occurred during the sessions, counselor leadership, relationships between counselors and students and the overall climate of sessions. These observations were conducted in the second year of our two-year evaluation, from November 2014 through the end of the 2014-15 academic year.

## V. RANDOMIZATION, RECRUITMENT, AND RETENTION PROCEDURES

### A. Randomization Procedures

During the summer preceding each study, the evaluation team identified eligible students in CPS schools using CPS administrative records from the previous academic year. In BAM 1, this included students who were in grades 7-10, in BAM 2 and BAM 2x2 this included students who were in grades 9 and 10, and in BAM expansion this mostly included students who were in grade 9. Following the approach used in Heller, et al. (2013), we first excluded those students who we thought were already too disengaged from school to attend regularly enough to benefit from a school-based program. This exclusion criterion was defined as having failed 75% or more of their classes the previous school year and having missed >60% of their enrolled school days in that year. We also excluded students with serious disabilities as designated by the CPS data.[10]

We then calculated a “risk index” that was a function of the number of prior-year course failures and unexcused absences, and being old for grade (interpreted as having been previously held back). Eligible students for the program were then ranked on the basis of this risk index. The research team determined the number of randomized students needed to utilize all available

[10] For our BAM 1 study, we use administrative data from the previous school year (AY2008-9) to create our study sample. Specifically, we receive enrollment data using administrative data from the previous school year. For our BAM 2, 2x2, and expansion samples, we receive enrollment data during the summer prior to the intervention, and match these enrollment files with administrative data from the first semester of the previous year (AY2012-13 for study 2 and study 2x2, and AY2013-14 for study expansion) to build our study sample.

program slots in a school and chose that number of students in descending order on the ranked risk list. The share of eligible students in each study sample varies across schools because of school-by-school variation in both program capacity and school size.

In practice, because of the scale of the experiments, in many schools we randomized all students who were not excluded based on their prior year course failures and absences. Essentially one can think of our study samples as pools of youth in distressed Chicago schools in the middle of the distribution for these schools, with both the left (lowest achieving) and right (highest achieving) tails trimmed.

In order to accommodate the varying program capacity within each school, our random assignment algorithm varied the probability of treatment assignment.[11] Since our randomization was carried out separately by school and grade for each study, we treat each school-grade combination as separate randomization blocks.[12]

In schools where too few students actually showed up in the fall to randomize, we identified new students entering the school (mostly during the first month of that school year) and randomly assigned them to treatment and control conditions. Specifically, during the school year, members of the evaluation team randomly assigned newly enrolled youth in each school using a spreadsheet pre-populated with treatment and control assignments for each new student added to the study sample. For these students the randomization block is defined by the school and the time period in which the youth was randomized.[13] All of our analyses below control for randomization-block fixed effects.

We include every student we randomized and for whom we have CPS data for in our analysis, including those who were assigned but subsequently left the pool of study schools.

We note that all randomization procedures were carried out in Stata for all four studies. Randomization into treatment and control groups was done by a member of the evaluation team using Stata. Balance among the assignment was verified by regressing treatment assignment against each of the pre-randomization characteristics we used to select our sample, while controlling for block-level fixed effects. Individual and joint significance was checked for each assignment group relative to the control group.

### B. Recruitment and Retention Procedures

After randomization occurred, the evaluation team sent lists of eligible students for the program to BAM counselors at each of the study sites. At the beginning of each study's school year, BAM counselors individually approached these eligible students and invited them to join the

[21] Our general rule was to randomize enough people into treatment groups to hit enrollment targets if we achieved a 75% take-up rate, and ideally to have a control group at least as big as the smallest treatment cell. In some schools because of the need to fill treatment slots, our control group was smaller than any of the treatment groups.

[12] We note one exception to this procedure in our first BAM study conducted in AY2009-10. In this study, each school, instead of school-grade, is treated as a separate randomization block.

[13] We note that we do not randomize new students entering the school in our first BAM study conducted in AY2009-10.

program. BAM counselors reviewed informed consent forms and processes in advance with the evaluation team, and followed up with both students and parents to directly answer any questions they had regarding participation in the program. Both parental consent and student assent were obtained from all program participants, and only those who had submitted both forms were able to participate in BAM programming. Students who were randomized into the control group were not approached for consent as they were only tracked through administrative data. Table 4 outlines the compliance rates and average number of sessions attended for each BAM study.

If we define attrition as the difference in year 1 and year 2 participation rates in our BAM studies that employed a two-year curriculum, we find that the BAM 2 sample has a 17% attrition rate, and the BAM 2x2 sample has a 13% attrition rate (specifically, 13% for those randomized only to receive BAM and 13% for those randomized to receive BAM and the math tutoring program).

## VI. ANALYSIS PLAN

### A. Estimating Main Program Impacts

We illustrate our approach to estimation in a simple regression framework. If *Y* is some social/behavioral outcome of interest, *S* is an indicator for assignment to Becoming a Man, *X* is a set of baseline youth characteristics or pre-randomization outcomes (included in the model to improve statistical power), and B is a set of randomization-block fixed effects, we would estimate the main intention to treat (ITT) effect as follows. (We omit student-level subscripts for convenience in what follows).

$$(1)\; Y = \pi_0 + \pi_1 S + \pi_2 X + \pi_3 B + \varepsilon_1$$

We initially estimate equation (1) with ordinary least squares, but for dichotomous dependent variables also re-estimate (2) using non-linear maximum likelihood models like probit and logit (although in practice the average marginal effects from the two approaches tend to be similar).

We also estimate the effects of participating in the program, or the effects of treatment on the treated (TOT), by imposing a linear functional form and using random assignment to different treatment conditions as instruments for actual participation (see Bloom 1984 and Angrist, Imbens and Rubin 1996, and for applications see Kling, Ludwig and Katz 2005, Kling, Liebman and Katz 2007, and Ludwig, et al. 2012). Whereas the ITT isolates the causal effect of being offered treatment, TOT estimates isolate the effect of the treatment for the subset of subjects who choose to participate. The TOT estimate is essentially the ratio of two experimental ITT effects: the ITT effect on the outcome of interest (*Y*) divided by the ITT effect on participation rates in the intervention being studied. This method recovers the TOT if assignment to treatment has no effect on outcomes for subjects who do not participate.[14, 15]

For all of our BAM evaluations, baseline demographic characteristics we use in regression models include: age, learning disability status, free/reduced lunch status, race, school-level randomization block, grade level at baseline, GPA at baseline, grades at baseline, school attendance, school disciplinary history (suspensions and incidents), and arrest history.

We can also estimate whether the average effects of BAM differ across the four studies by pooling together the student-level data. To test the null hypothesis of no difference in the average ITT effect across BAM studies, we create study-specific fixed effects and include them together with interactions of the study fixed effects and treatment assignment variable in equation (1) above. We then carry out an F-test that the interaction terms of study with treatment assignment are jointly equal to zero.

Since the compliance or "take-up" rate is somewhat different across BAM studies, it is possible the ITT effect could be different even if the average effect on participants is the same. To explore that possibility, we also re-estimate a version of our pooled sample analysis but now using interactions of study fixed effect and treatment assignment as instrumental variables for interactions of study fixed effect and actual program participation. We can then carry out an F-test for whether the interactions between study and participation are jointly zero.

### B. Understanding Mediating Factors

Mediating factors are those that lie in the causal chain in between BAM program participation and the outcome; that is, BAM → M → Y. To better understand how BAM works, we use two-stage least squares to estimate the system of equations given by (2) and (3) below. In the first-stage equation, we use interactions of the randomization-block fixed effects and treatment assignment as instruments for different candidate mediators, while the second-stage equation replaces the actual with the predicted values for the candidate mediators to understand their relationship to the outcome of interest:

$$(2)\ M = \alpha_0 + \alpha_1 S + \alpha_2 X + \alpha_3 B + \alpha_4 B * S + \varepsilon_2$$

$$(3)\ Y = \pi\beta_0 + \beta_1 M + \beta_2 X + \beta_3 B + \varepsilon_3$$

This follows the approach from Kling, Liebman and Katz (2007) and exploits the variation across randomization blocks in the degree to which program participation changes candidate

[14] If no controls participate in the program, then our instrumental variables estimate identifies the average effect of everyone who is "treated" (participates in the program), the TOT. If some control group members wind up receiving program services, which in a complicated real-world setting like CPS could potentially happen to some small degree, then our IV estimates are still valid, they just estimate a slightly different parameter – the local average treatment effect (LATE) for subjects who participated because they were selected to be in the treatment group but would not have participated if they had been in the control group. This group is called "compliers" in the typology of Angrist, Imbens, and Rubin 1996. We begin our analysis with models using school-level fixed effects.

[15] For example, if treatment assignment results in a 5% increase in outcome Y, and the participation rate is 50%, our TOT estimate becomes 0.05/0.5, or a 10% increase.

mediators. The identifying assumption is the only reason for variation across randomization blocks in BAM impacts is variation in how BAM affects the specified mediator of interest.

One challenge in estimating equations (2)-(3) is that some randomization blocks have relatively few students. So we also re-estimate these equations now using interactions of treatment assignment with school, rather than randomization block, fixed effects.

A different challenge stems from the fact that we have a limited number of schools and randomization blocks in our data, combined with a substantial amount of variability in our outcomes of interest. This means that in practice we wind up having somewhat limited statistical power to understand mediation pathways, as we discuss further below.

### C. Understanding Moderators

Finally, we also seek to better understand how different baseline characteristics of students and schools (the Xs) moderate the effect of BAM on outcomes. This is especially important since descriptively the population served in later BAM RCTs had a lower baseline arrest rate and was more likely to be Latinx than the first two studies. For this purpose we now re-estimate equation (1) including an interaction of a given baseline characteristic with treatment assignment, as in equation (4). We should note that while treatment was randomly assigned, the baseline student- and school-level characteristics are obviously not. So the estimated moderation coefficient $\delta_4$ should not be given a causal interpretation in the sense of assuming that some exogenous policy manipulation of a given X would change the treatment response by that exact amount. It is possible that the baseline characteristic that we happen to include in the estimating equation is correlated with some other omitted moderator. Nonetheless, this analysis can help us understand how differences across BAM study samples and schools might help explain variation across studies in estimated overall average program impacts.

$$(4)\ Y = \delta_0 + \delta_1 S + \delta_2 X + \delta_3 B + \delta_4 S * X + \varepsilon_4$$

One limitation of the approach described above is that more than one baseline characteristic may moderate BAM impacts on youth outcomes. In principle, it is even possible that there may be non-linearities in how the baseline variables moderate the outcome (say, prior student achievement test scores moderate BAM impacts only when they are above or below some level), or interactions in how multiple baseline characteristics together moderate BAM (for example, if having low prior test scores only affects BAM impacts if the student attends school in a very high-crime neighborhood). This creates a challenge for standard econometric methods in the sense that including every possible non-linearity and interaction of our baseline variables in equation (4), and their interactions with treatment assignment, can lead to a regression equation with a set of right hand side variables that rapidly approaches our total number of observations.

To address this challenge, we use the machine learning approach of Athey and Imbens (2017). The basic idea is to treat the problem of modeling the structure of the moderators (or, equivalently, the structure of the heterogeneous treatment effects, or HTEs) as a prediction problem. A simple version of this would use a decision tree to consider each possible split on each baseline variable in the dataset, to divide the data up into two groups that are as similar within groups as possible, and different across groups as possible, in terms of their average estimated BAM effect. The algorithm then splits those two groups again; since the baseline variable that is selected for the next split on the left need not be the same as on the right, and since the same variable can be used at different splits in the tree (that is, different value cutoffs for the same variable can be used for multiple splits), the decision tree allows for a great deal of interactivity and non-linearity in moderating how the baseline characteristics moderate program impacts. The decision tree recursively splits those two groups until some stopping rule is reached. The actual implementation of this approach uses a more sophisticated variant of decision trees that averages multiple trees together to avoid over-fitting the data, cross-validation to select the model with the optimal level of complexity or expressiveness, and evaluates its performance in a separate randomly-selected "hold out" set of observations different from the set of observations used to build the initial model. For statistical inference we use the bootstrap.

## VII. MISSING DATA

We distinguish between two types of missing data in our study: missing data from the baseline variables used as control variables, and missing data in outcome variables.

In general, youth in study 1 have lower levels of missing baseline information for grades, which is a result of the study team's use of the prior year enrollment file to select youth for randomization. In BAM 2, BAM 2x2, and BAM expansion studies, the number of youth missing grade data is higher due to the use of summer enrollment files, which include new youth who transferred to Chicago Public Schools after the school year ended. See Table 5 for missingness of attendance and grades by study.

From the perspective of valid statistical inference on program impacts, what matters is not so much the overall level of missingness on outcome data but rather the degree to which it differs between the randomized treatment and control groups. In BAM 1 there is some evidence of a statistically significant difference in missing attendance data between groups, but otherwise there is no evidence of differences within and across studies. Test of differences is carried out by regressing an indicator for missing data against an indicator for treatment assignment, controlling for randomization blocked fixed effects. We then assess the significance of the treatment assignment coefficient.

For youth missing baseline variables, we created new variables with zeros imputed for any missing observations. We use these along with indicator variables flagging students who had a given baseline variable imputed, into all models.

Missing data in the outcome variables increases each of the post-randomization period. See Table 6 for missingness of outcome variables by study.

Using the same test described above for missing baseline data, we see some evidence of a difference in the proportion of treatment and control youth missing grade data in the BAM expansion sample's first year. There is no further evidence of differences within or across studies. By construction there is no missing data in the baseline or outcome arrest variables, nor in the third element of the school index, which measures whether or not youth were still enrolled at the end of a given school year.

For youth missing outcome data for grades, attendance, or both, we impute the treatment or control mean. Heller, et al. (2013) describes this approach as having "the advantage of using all available information and having a straightforward substantive interpretation: it is equivalent to estimating the treatment effect on each individual component of the index (in standardized form) using only observations with non-missing observations, and then averaging the component-specific estimates (Kling, Liebman and Katz 2007, Anderson 2008, Schochet, Burghardt and McConnell 2008)." This assumes that missing grade or attendance data is not related to any observable or unobservable characteristics of the youth, nor to any outcome data. In other words, the data are assumed to be missing completely at random (MCAR).

## VIII. RESULTS

### A. Impacts for each individual BAM RCT

Tables 7-10 summarize the results of each of the studies separately. In sum, we find that our BAM 1 and BAM 2 studies show consistently positive effects, while the BAM expansion study shows more mixed effects, and the BAM 2x2 study shows signs of null-to-adverse effects.

To compare the results across studies requires making some decisions about when to measure the outcomes. Some studies involved a 1-year curriculum, while others had a 2-year curriculum, and for one of our studies, there was a 2-year curriculum but we only have data through the end of the first program year. Thus there is no obviously perfect way to compare outcomes. We err on the side of showing the reader more rather than less of the data, and present results for each study in what we believe are four different reasonable ways:

- The effect measured at the end of the first program year, which will then have the drawback of including data on the BAM 2, BAM 2x2, and BAM expansion measured essentially during the middle of the two-year program period;
- The effect measured at the end of the second program year, which has the drawback of excluding data from the BAM 1 study and BAM expansion;
- The average effect calculated over all program years for which we have data;
- And the effect measured at the end of program completion period, which excludes data from the BAM expansion study.

Tables 7 and 8 show that there are large beneficial impacts in our BAM 1 and BAM 2 studies, reported on in Heller et al. (2017), which are quite similar in magnitude. We find a statistically significant impact on school engagement by 0.14 standard deviations (which we estimate will translate into an increase in graduation rates of 9.7%) and a decrease in violent crime arrests by 45% in our BAM 1 study, and improvements in school engagement and violent crime arrests that vary in impact and statistical significance for our BAM 2 study depending on the model used. We also see reductions in all arrests and "other" arrests. In contrast, the estimates for the BAM expansion RCT are mixed: the results are generally not statistically significant, but they are large in magnitude, with an adverse effect on school engagement and sizable proportional reductions in every arrest category except for drug offenses (Table 10). And for the BAM 2x2 study (Table 9) there are if anything consistently adverse effects on both school engagement and arrests.

Figures 1-3 present the plausible range of coefficient estimates for each of the main study outcomes using a slightly different methodology—sampling using a bootstrap approach.[16] These figures show that while statistical significance varies across studies and outcomes, there is general consistency in the magnitude of estimated coefficients in the direction of beneficial impacts. While the results are generally consistent across studies, there are still some outliers—notably for the BAM 2x2 study—that require further examination.

An example of this can be seen with Figure 2, which presents results on arrests to youth for all crimes. Whether we measure outcomes averaged over all program years, or focus on outcomes measured at the end of the program participation period, effects of BAM participation are typically in the range of reductions in arrests of between -9% to -34%. The exception is the BAM 2x2 study, where we see sizable adverse effects – but also a sizable confidence interval. The magnitude is itself somewhat sensitive to when we measure program impacts (comparing across estimates for the "average effect" and when "measured at program completion" in Table 9), which highlights that the estimate itself is being driven in part by a few schools that have large, outlier impacts in selected years.

### B. Testing for differences in average impacts across studies

While the point estimates show variability across several of the BAM RCTs, the confidence intervals around these estimates are often quite sizable. In addition to understanding the treatment effects of BAM *within* studies, we want to understand whether the treatment effects *across* studies are statistically distinguishable from one another. In this section we carry out formal statistical tests of the null hypothesis that the effects are the same across studies.

We pool student-level data from each study and run our previously specified ITT model with the addition of study-specific fixed effects and interactions of those study effects with treatment assignment. To test whether the ITT effect of BAM differs across studies, we conduct an F-test for the joint null hypothesis that the interaction terms of study and treatment assignment are all equal to zero.[17] In order to account for the risk of type I error in multiple hypothesis testing, we adjust our estimates using the Familywise Error Rate (FWER) procedure suggested by Westfall and Young (1993). Table 11 summarizes the results of these tests.

Panel 1 shows the p-value (unadjusted for the number of statistical tests we are conducting) for a given F-test of the null hypothesis that the effect of BAM on the measure (e.g. "All arrests") shown in a given column is the same across all four BAM studies within the outcome window specified in a given row (e.g. "Average effect"). For example, the first cell shows that if we look at the estimated effect of BAM on school engagement as measured at the end of the first-year of

[16] For these figures, we use a bootstrap approach, sampling with replacement to generate a thousand estimates for each model of interest. We take the 2.5th and 97.5th percentile estimates as the confidence interval, and the 50th percentile as the median listed estimate.

[17] Since we are interested in the difference between program effects and not the magnitude of the underlying effects, we only conduct this exercise for our ITT models. However, we would expect the results to be mirrored if we carried out a similar test for difference in the TOT effect across studies by using interactions of study and treatment assignment as instruments of program participation and study.

each BAM study, we can reject the null that the effects are the same across all four studies at the 5% level since p=.034 for this test.

Panels 2-10 show the p-values for these same tests, but adjusted for multiple hypothesis testing. In each panel, we present adjusted p-values using a different choice for the group of outcomes being tested and, consequently, the number of statistical tests being conducted. In Panel 2, we present the full set of outcomes (all six measures defined across each of the four outcome windows) as a family. In Panel 3, we present the same set from Panel 2, but excluding the measure of "all arrests" since this measure is fully represented by the sum of the four arrest sub-types (violent, property, drug, and other). Panels 4-5 treat the set of outcomes measured as the average of outcomes across all program years ("Average effect") as a family, both with and without the "all arrests" measure, respectively. Panels 6-7 present the same set of outcomes as Panels 4-5, but measured at the end of each study ("Program completion effect"). Finally, Panels 8-10 present each definition of our three primary outcomes ("School engagement", "All arrests", and "Violent arrests") as a family, respectively.

While several of the unadjusted tests are significant, when adjusting for multiple hypothesis testing using different families of outcomes we consistently fail to reject the null that treatment effects are the same across studies. One challenge with multiple testing procedures is that they require grouping outcomes into families and, as in our case, there are often multiple defensible ways to do this (for example, defining families based on the behavioral outcomes, or defining families based on the point in time when the outcome is measured). When we use different definitions of families here, the p-values vary somewhat and sometimes approach, but never reach, statistical significance at the 5% level (in one case, the Average effect on school engagement in Panel 8, the p-value reaches significance at the 10% level).

### C. Test of Mediation

While we cannot reject the null hypothesis that the program impacts are the same across studies, the size of our confidence intervals allows for substantial variation across RCTs in average effects. To the extent to which there are differences across BAM RCTs in program effects, we begin to try to understand them here by exploring whether BAM impacts on candidate mediators line up with the pattern of BAM impacts on outcomes.

### D. Test of Moderators

We next turn to understanding which, if any, of the baseline student- or school-level variables we have data for might moderate average program impacts (to the extent to which there might be differences in average effects that we do not have statistical power to detect, given the large differences in point estimates across some of our studies but also the wide confidence intervals). Again, we think this is an important exercise given the variability in students served, implementation fidelity, school settings, and neighborhood contexts in each of the study periods.

For a baseline variable to moderate impacts across BAM RCTs, three conditions are required:

The mean values of the moderator must differ systematically across studies in a way that has a similar rank-ordering (positive or negative) to program impacts on outcomes
We should also see a statistically significant interaction between treatment assignment and that moderator in a model that has one of our outcomes as the dependent variable
We should see both of the previous patterns for each of our key outcomes (schooling engagement, total arrests, violent crime arrests)

We first examine each of the candidate moderators listed in Table 12 one at a time to see if they pass the conditions described above. One of the only baseline characteristics that meets both required conditions is the *race / ethnicity of the student*. Specifically, African-American students seem to benefit more on average than do Hispanics, and the share of program participants who are African-American declines in later BAM RCTs.

However, race / ethnicity does not seem to *fully* explain any variability in program impacts across RCTs. We can see this in Figures 4-6 where we show separate BAM impacts for African-Americans and Hispanics in each BAM RCT over time. There is a smaller decline in the point estimates across RCTs for African-Americans than Hispanics, but still a difference across studies in point estimates even for African-American youth. (We note again the sizable standard errors here, and so focus on point estimates for now).

A different way to explore this issue of moderation is by considering how the full set of baseline characteristics moderate program impacts, alone or in combination, using the machine learning method of Athey and Imbens (2017). After all, if a single baseline characteristic explains some of the variation across studies in average program impacts, perhaps considering multiple baseline characteristics at the same time might explain more of the impact variability?

We first estimate the heterogeneous treatment effects (HTEs) using this method applied to the BAM 1 RCT, where we see large impacts. We then apply these HTE estimates to subsequent RCTs and calculate the implied average effects of each subsequent BAM study if the effect for a given “type” of student had stayed the same as it was in BAM 1, but just the composition of the study in terms of student “types” had changed. This exercise is in the spirit of a Blinder (1973) or Oaxaca (1973) decomposition. If program effects were constant and all of the variation in program impacts were just due to changes in student composition, we would expect the simulated average impact in this case to be small or zero (since the subsequent studies would be over-represented by the “types” of students who do not benefit much). If, on the other hand, the main driver of the difference in program effects was some change in BAM average effectiveness over time, the simulated effect would be similar to what we see in BAM 1 (that is, the change in composition of types of students would have negligible effects). And indeed this is what we indeed see in Table 13 and Figure 7.

The only moderator that seems to consistently pass this test is the number of youth citywide receiving BAM during a given academic year, as shown in Figure 8 Since there is some overlap in the years in which some studies were carried out, we now pool studies by year and plot the average effects for all studies carried out in a given year against the number of youth who received BAM anywhere in Chicago during that year (whether those youth were in one of our

BAM RCTs or not, since total number of youth in BAM citywide seems like the most relevant measure of program scale).

**Table 1: Select BAM Activities**

| Activity Category | Example Activities |
|---|---|
| Reflective/ Introspective | **Check-Ins:** Students talk to each other about what they are doing well and areas where they still need to improve. Students must listen patiently while someone else discusses their attributes. |
| Immersive/ Experiential | **The Fist:** Students are told to get an object from a partner. Many try to use force. The counselor asks questions to highlight how their partners were willing to give up the object if they calmly requested it. |
| | **Plates:** Students reflect on what it has taken to successfully complete group missions and write those attributes on a plate. The plates are placed on the floor, and students must cross the floor by using the plates. However, if no one is standing on a plate, then it is removed (making the task more difficult). |
| | **Trust Walk:** Students follow group leaders around the school silently and without disrupting the school. They are told that with freedom comes responsibility. |
| | **Focus Mitt Drill:** Students punch focus mitts for an extended period. |
| | **Human Knot:** Students stand in a circle and grab the hands of someone standing across from them. They must then untangle themselves without letting go. |
| Role Playing | **$10 Role Play:** Students role play a student borrowing money and then never paying it back. |
| | **High School Day:** Students do a role-play where a student and administrator have a confrontation. They act out the conflict with “out of control” and “in control” anger expressions. |
| | **Our Story Of What Happened:** Students imagine a conflict and discuss why the conflict came about. They examine thinking distortions that might have made the conflict worse. |
| | **Rudy:** Students watch and discuss the movie Rudy. Before beginning the movie, the counselor holds up two dollars and asks who wants the money. Even as students raise their hand, he keeps asking who wants it until someone simply takes it from him. He explains that we often overlook opportunities, but the student who took the money saw it as an opportunity and took a chance. |

|  |  |
|---|---|
| Stories & Discussion | **The Boy Who Cried Wolf:** Students listen to and discuss the story where one day a boy pretends that he is being attacked by a wolf. He is amused by how his town responds to this prank. So when he feels bored on another day, he does it again. And again. He promises to stop playing around, but when he feels bored he can't help but do it again. In the end, when he is actually attacked by a wolf, no one responds to his pleas for help. |
|  | **Miracle:** Students watch and discuss the film Miracle about the U.S. men's hockey team. |
| Skill- building | **Cognitive Thought Replacement:** Students learn how to recognize negative thoughts that arise and how to replace them. It is not necessary to replace negative thoughts with positive thoughts, but rather to instead focus on what can be done to control the situation that is leading to the negative thought. |
|  | **Manhood Questions and Rites of Passage:** Students discuss the key moments when boys become men and various rites of passage that exist. |
|  | **Positive Anger Expression:** Students are taught about how to express anger in a controlled way. |

## Table 2: Overview of all studies

| ***Study*** | 2009-10 | 2010-11 | 2011-12 | 2012-13 | 2013-14 | 2014-15 | 2015-16 |
|---|---|---|---|---|---|---|---|
| **BAM 1**<br>**(Study 1)**<br>***(18 schools)*** | N= 2,740 | | | | | | |
| **BAM 2**<br>**(Study 2)**<br>***(9 schools)*** | | | | | N=2,064 | ▒ | |
| **BAM 2x2**<br>**(Study 3)**<br>**(*12 schools)*** | | | | | N=2,633 | ▒ | |
| **BAM**<br>**Expansion**<br>**(Study 4)**<br>**(*21* schools)** | | | | | | N=2,367 | ▓ |

## Table 3: Becoming a Man Studies – Baseline Characteristics

| | Study 1 | | Study 2 | | Study 3 | | Study 4 | |
|---|---|---|---|---|---|---|---|---|
| | Control | Treatment | Control | Treatment | Control | Treatment | Control | Treatment |
| **Number of Students** | 1267 | 1473 | 1048 | 1016 | 1453 | 1180 | 1282 | 1085 |
| **Demographics** | | | | | | | | |
| Black | 0.720 | 0.688 | 0.698 | 0.683 | 0.456 | 0.479 | 0.632 | 0.610 |
| Hispanic | 0.276 | 0.307 | 0.275 | 0.300 | 0.493 | 0.464 | 0.325 | 0.357 |
| English language learner | | | 0.050 | 0.053 | 0.116 | 0.116 | 0.079 | 0.105 |
| Age | 15.700 | 15.512 | 14.845 | 14.910 | 14.802 | 14.811 | 14.484 | 14.457 |
| Free lunch recipient | 0.899 | 0.910 | 0.845 | 0.834 | 0.866 | 0.869 | 0.885 | 0.895 |
| Learning disability | 0.198 | 0.186 | 0.168 | 0.164 | 0.178 | 0.162 | 0.178 | 0.163 |
| **Schooling** | | | | | | | | |
| Grade 9 | 0.455 | 0.450 | 0.597 | 0.543 | 0.552 | 0.580 | 0.864 | 0.899 |
| Grade 10 | 0.493 | 0.445 | 0.395 | 0.448 | 0.434 | 0.410 | 0.121 | 0.089 |
| GPA | 1.679 | 1.734 | 2.111 | 2.157 | 2.120 | 2.068 | 2.317 | 2.260 |
| **Crime** | | | | | | | | |
| Any arrests at baseline | 0.369 | 0.346 | 0.230 | 0.232 | 0.182 | 0.188 | 0.186 | 0.186 |
| Number of baseline arrests for: | | | | | | | | |
| Violent offenses | 0.353 | 0.348 | 0.185 | 0.184 | 0.138 | 0.132 | 0.139 | 0.153 |
| Property offenses | 0.206 | 0.189 | 0.138 | 0.129 | 0.081 | 0.090 | 0.089 | 0.102 |
| Drug offenses | 0.168 | 0.177 | 0.111 | 0.144 | 0.069 | 0.064 | 0.073 | 0.061 |
| Other offenses | 0.449 | 0.470 | 0.290 | 0.321 | 0.208 | 0.238 | 0.234 | 0.234 |

Note: P-value on F-test of treatment-control comparison for all baseline characteristics: Study 1: p=0.668; Study 2: p=0.499; Study 3: p=0.882; Study 4: p=0.717. Joint significance tests for equality of all baseline characteristics use only non-missing data. Grade level measured at start of study. Asterisks indicate statistical significance of pairwise treatment-control comparison for a given baseline characteristic controlling for randomization block fixed effects with heteroscedasticity- robust standard errors. Data from Chicago Public Schools administrative data and Chicago Police Department arrest records. Means calculated using non- missing observations for each variable. Pre-program arrests are arrests prior to start of program school year. GPA is measured on a 0-4 scale. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

**Table 4: Compliance and Attendance for Youth Assigned to BAM**

| Study | Compliance rate (attend >0 sessions) | Average sessions attended among all assigned to BAM | Average sessions among those with >0 sessions attended |
|---|---|---|---|
| BAM 1 | 49% | 5.2 | 11.0 |
| BAM 2 | 52% | 12.7 | 24.1 |
| BAM 2x2 | 50% | 14.4 | 28.6 |
| BAM Expansion | 31% | | |

**Table 5: Missing Baseline Variables by Study**

| | Missing prior grades | Missing prior attendance |
|---|---|---|
| BAM 1 | 6% | 5% * |
| BAM 2 | 14% | 4% |
| BAM 2x2 | 7% | 3% |
| BAM Expansion | 9% | 4% |

Significance is indicated by *** if p<0.01, by ** if p<0.05, and by * if p<0.1

**Table 6: Missing Outcome Variables by Study**

| | Missing grades, year 1 | Missing grades, year 2 | Missing attendance, year 1 | Missing attendance, year 2 |
|---|---|---|---|---|
| BAM 1 | 10% | 33% | 3% | 17% |
| BAM 2 | 27% | 39% | 7% | 17% |
| BAM 2x2 | 15% | 27% | 5% | 14% |
| BAM Expansion | 23%* | 32% | 8% | 16% |

Significance is indicated by *** if p<0.01, by ** if p<0.05, and by * if p<0.1

## Table 7: Becoming a Man Study 1 – Effect on Youth Outcomes

| Model | N | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean |
|---|---|---|---|---|---|---|---|---|---|
| | | | **School Engagement Index** | | | | **All Arrests** | | |
| First-year effect | 2740 | 0.0000 | 0.0584*** | 0.1400*** | 0.2184 | 0.6993 | -0.0744 | -0.1784 | 0.6640 |
| | | | (0.0214) | (0.0508) | | | (0.0472) | (0.1123) | |
| Average effect | 2740 | 0.0000 | 0.0584*** | 0.1400*** | 0.2184 | 0.6993 | -0.0744 | -0.1784 | 0.6640 |
| | | | (0.0214) | (0.0508) | | | (0.0472) | (0.1123) | |
| Measured at program completion | 2740 | 0.0000 | 0.0584*** | 0.1400*** | 0.2184 | 0.6993 | -0.0744 | -0.1784 | 0.6640 |
| | | | (0.0214) | (0.0508) | | | (0.0472) | (0.1123) | |
| | | | **Violent Arrests** | | | | **Property Arrests** | | |
| First-year effect | 2740 | 0.1665 | -0.0346** | -0.0829** | 0.1860 | 0.0766 | 0.0075 | 0.0180 | 0.0599 |
| | | | (0.0166) | (0.0395) | | | (0.0128) | (0.0305) | |
| Average effect | 2740 | 0.1665 | -0.0346** | -0.0829** | 0.1860 | 0.0766 | 0.0075 | 0.0180 | 0.0599 |
| | | | (0.0166) | (0.0395) | | | (0.0128) | (0.0305) | |
| Measured at program completion | 2740 | 0.1665 | -0.0346** | -0.0829** | 0.1860 | 0.0766 | 0.0075 | 0.0180 | 0.0599 |
| | | | (0.0166) | (0.0395) | | | (0.0128) | (0.0305) | |
| | | | **Drug Arrests** | | | | **Other Arrests** | | |
| First-year effect | 2740 | 0.1507 | -0.0001 | -0.0003 | 0.1003 | 0.3054 | -0.0472* | -0.1132* | 0.3179 |
| | | | (0.0183) | (0.0434) | | | (0.0278) | (0.0663) | |
| Average effect | 2740 | 0.1507 | -0.0001 | -0.0003 | 0.1003 | 0.3054 | -0.0472* | -0.1132* | 0.3179 |
| | | | (0.0183) | (0.0434) | | | (0.0278) | (0.0663) | |
| Measured at program completion | 2740 | 0.1507 | -0.0001 | -0.0003 | 0.1003 | 0.3054 | -0.0472* | -0.1132* | 0.3179 |
| | | | (0.0183) | (0.0434) | | | (0.0278) | (0.0663) | |

Note: Table presents estimates for the effect of BAM in Study 1. The first-year, average, and program completion effects measure outcomes from the first year. The second-year effect is omitted because there was no second study year in Study 1. Participation for the IV is attending at least one session during the first year. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroscedasticity-robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

## Table 8: Becoming a Man Study 2 – Effect on Youth Outcomes

| Model | N | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean |
|---|---|---|---|---|---|---|---|---|---|
| | | | **School Engagement Index** | | | | **All Arrests** | | |
| First-year effect | 2064 | 0.0000 | 0.0148 | 0.0297 | 0.2050 | 0.5954 | -0.0921* | -0.1844* | 0.6455 |
| | | | (0.0245) | (0.0482) | | | (0.0497) | (0.0984) | |
| Second-year effect | 1872 | 0.0000 | 0.0543** | 0.1716** | 0.1887 | 0.6197 | -0.0550 | -0.1740 | 0.5040 |
| | | | (0.0269) | (0.0833) | | | (0.0581) | (0.1812) | |
| Average effect | 2064 | -0.0140 | 0.0349* | 0.0692* | 0.1241 | 0.5825 | -0.0704 | -0.1394* | 0.6185 |
| | | | (0.0211) | (0.0409) | | | (0.0429) | (0.0841) | |
| Measured at program completion | 1872 | 0.0000 | 0.0543** | 0.1110** | 0.0939 | 0.6197 | -0.0550 | -0.1126 | 0.6628 |
| | | | (0.0269) | (0.0541) | | | (0.0581) | (0.1174) | |
| | | | **Violent Arrests** | | | | **Property Arrests** | | |
| First-year effect | 2064 | 0.1193 | -0.0223 | -0.0447 | 0.1294 | 0.0725 | -0.0059 | -0.0119 | 0.0710 |
| | | | (0.0158) | (0.0313) | | | (0.0126) | (0.0248) | |
| Second-year effect | 1872 | 0.1165 | -0.0280 | -0.0884 | 0.1817 | 0.0716 | -0.0034 | -0.0108 | 0.0608 |
| | | | (0.0191) | (0.0599) | | | (0.0141) | (0.0440) | |
| Average effect | 2064 | 0.1150 | -0.0244* | -0.0483* | 0.1374 | 0.0687 | -0.0031 | -0.0062 | 0.0747 |
| | | | (0.0131) | (0.0256) | | | (0.0101) | (0.0197) | |
| Measured at program completion | 1872 | 0.1165 | -0.0280 | -0.0572 | 0.1598 | 0.0716 | -0.0034 | -0.0070 | 0.0899 |
| | | | (0.0191) | (0.0386) | | | (0.0141) | (0.0285) | |
| | | | **Drug Arrests** | | | | **Other Arrests** | | |
| First-year effect | 2064 | 0.1269 | -0.0178 | -0.0356 | 0.1734 | 0.2767 | -0.0460 | -0.0921 | 0.2717 |
| | | | (0.0236) | (0.0467) | | | (0.0289) | (0.0572) | |
| Second-year effect | 1872 | 0.1474 | 0.0017 | 0.0053 | 0.0781 | 0.2842 | -0.0253 | -0.0801 | 0.1834 |
| | | | (0.0250) | (0.0778) | | | (0.0327) | (0.1018) | |
| Average effect | 2064 | 0.1307 | -0.0081 | -0.0160 | 0.1550 | 0.2681 | -0.0348 | -0.0689 | 0.2515 |
| | | | (0.0184) | (0.0359) | | | (0.0238) | (0.0466) | |
| Measured at program completion | 1872 | 0.1474 | 0.0017 | 0.0034 | 0.1494 | 0.2842 | -0.0253 | -0.0518 | 0.2636 |
| | | | (0.0250) | (0.0504) | | | (0.0327) | (0.0659) | |

Note: Table presents estimates for the effect of BAM in Study 2. The first-year effect measures outcomes from the first year; participation for the IV is attending at least one session during the first year. The second-year effect measures outcomes from the second year; participation for the IV is attending at least one session during the second year. The average effect measures the mean of outcomes from the first year and second years; for Johnson the average effect is just the first year; participation for the IV is attending at least one session during either year. The program completion effect measures outcomes from the second year; participation for the IV is attending at least one session during either year. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroscedasticity-robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

## Table 9: Becoming a Man Study 3 – Effect on Youth Outcomes

| Model | N | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean |
|---|---|---|---|---|---|---|---|---|---|
| | | | School Engagement Index | | | | All Arrests | | |
| First-year effect | 2302 | 0.0000 | -0.0033 | -0.0070 | 0.2792 | 0.3089 | 0.0305 | 0.0641 | 0.1754 |
| | | | (0.0255) | (0.0529) | | | (0.0362) | (0.0751) | |
| Second-year effect | 2302 | 0.0000 | -0.0144 | -0.0412 | 0.3755 | 0.3012 | 0.0663* | 0.1902* | 0.0192 |
| | | | (0.0241) | (0.0686) | | | (0.0386) | (0.1095) | |
| Average effect | 2302 | 0.0000 | -0.0089 | -0.0185 | 0.2256 | 0.3050 | 0.0484 | 0.1011* | 0.1776 |
| | | | (0.0217) | (0.0447) | | | (0.0297) | (0.0612) | |
| Measured at program completion | 2302 | 0.0000 | -0.0144 | -0.0300 | 0.1754 | 0.3012 | 0.0663* | 0.1385* | 0.1717 |
| | | | (0.0241) | (0.0498) | | | (0.0386) | (0.0795) | |
| | | | Violent Arrests | | | | Property Arrests | | |
| First-year effect | 2302 | 0.0687 | -0.0124 | -0.0260 | 0.0668 | 0.0448 | 0.0091 | 0.0191 | 0.0070 |
| | | | (0.0122) | (0.0253) | | | (0.0107) | (0.0223) | |
| Second-year effect | 2302 | 0.0525 | 0.0116 | 0.0334 | 0.0098 | 0.0355 | 0.0097 | 0.0278 | -0.0012 |
| | | | (0.0116) | (0.0328) | | | (0.0101) | (0.0285) | |
| Average effect | 2302 | 0.0606 | -0.0004 | -0.0008 | 0.0452 | 0.0402 | 0.0094 | 0.0196 | 0.0217 |
| | | | (0.0093) | (0.0191) | | | (0.0081) | (0.0166) | |
| Measured at program completion | 2302 | 0.0525 | 0.0116 | 0.0243 | 0.0240 | 0.0355 | 0.0097 | 0.0202 | 0.0323 |
| | | | (0.0116) | (0.0239) | | | (0.0101) | (0.0207) | |
| | | | Drug Arrests | | | | Other Arrests | | |
| First-year effect | 2302 | 0.0548 | 0.0071 | 0.0149 | 0.0374 | 0.1405 | 0.0267 | 0.0561 | 0.0642 |
| | | | (0.0151) | (0.0313) | | | (0.0212) | (0.0440) | |
| Second-year effect | 2302 | 0.0718 | 0.0146 | 0.0418 | 0.0389 | 0.1413 | 0.0304 | 0.0872 | -0.0283 |
| | | | (0.0170) | (0.0482) | | | (0.0220) | (0.0624) | |
| Average effect | 2302 | 0.0633 | 0.0108 | 0.0226 | 0.0502 | 0.1409 | 0.0285* | 0.0596* | 0.0605 |
| | | | (0.0124) | (0.0255) | | | (0.0173) | (0.0357) | |
| Measured at program completion | 2302 | 0.0718 | 0.0146 | 0.0304 | 0.0612 | 0.1413 | 0.0304 | 0.0635 | 0.0542 |
| | | | (0.0170) | (0.0351) | | | (0.0220) | (0.0453) | |

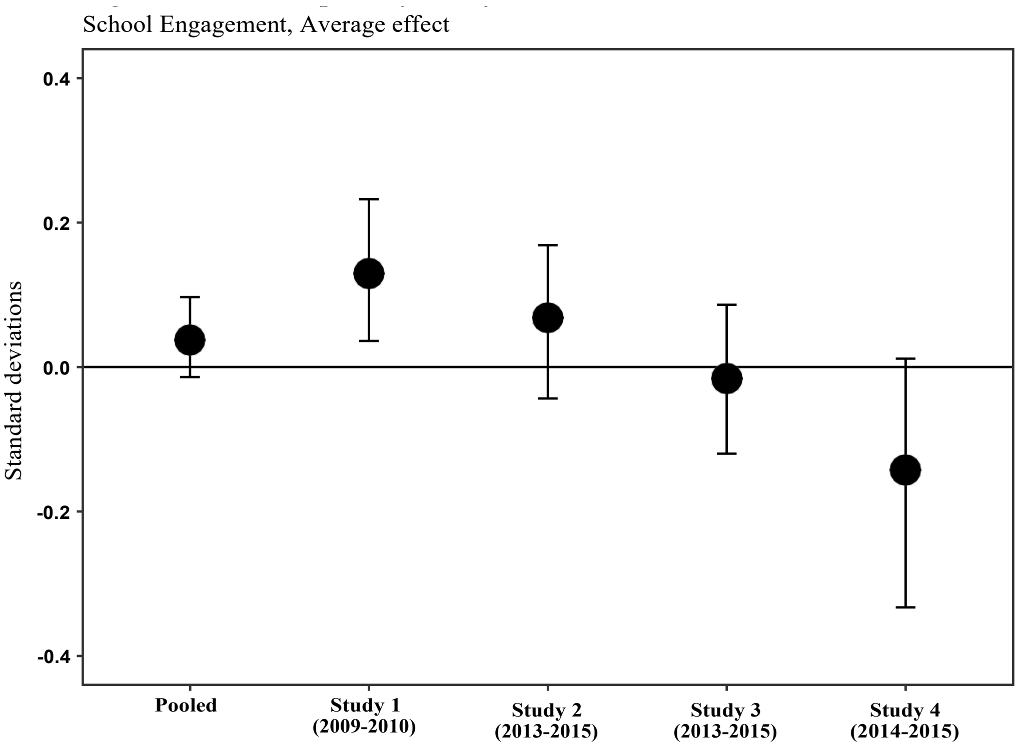
Note: Table presents estimates for the effect of BAM in Study 3. The first-year effect measures outcomes from the first year; participation for the IV is attending at least one session during the first year. The second-year effect measures outcomes from the second year; participation for the IV is attending at least one session during the second year. The average effect measures the mean of outcomes from the first year and second years; participation for the IV is attending at least one session during either year. The program completion effect measures outcomes from the second year; participation for the IV is attending at least one session during either year. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroscedasticity-robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

## Table 10: Becoming a Man Study 4 – Effect on Youth Outcomes

| Model | N | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean | Control Mean | Intention to Treat | Effect of Participation (IV) | Control Complier Mean |
|---|---|---|---|---|---|---|---|---|---|
| | | | **School Engagement Index** | | | | **All Arrests** | | |
| First-year effect | 2201 | 0.0000 | -0.0522** | -0.1742* | 0.4375 | 0.4489 | -0.0409 | -0.1364 | 0.4491 |
| | | | (0.0266) | (0.0892) | | | (0.0418) | (0.1370) | |
| Average effect | 2201 | 0.0000 | -0.0522** | -0.1742* | 0.4375 | 0.4489 | -0.0409 | -0.1364 | 0.4491 |
| | | | (0.0266) | (0.0892) | | | (0.0418) | (0.1370) | |
| | | | **Violent Arrests** | | | | **Property Arrests** | | |
| First-year effect | 2201 | 0.0956 | -0.0101 | -0.0336 | 0.1197 | 0.0640 | -0.0050 | -0.0168 | 0.0438 |
| | | | (0.0141) | (0.0462) | | | (0.0140) | (0.0460) | |
| Average effect | 2201 | 0.0956 | -0.0101 | -0.0336 | 0.1197 | 0.0640 | -0.0050 | -0.0168 | 0.0438 |
| | | | (0.0141) | (0.0462) | | | (0.0140) | (0.0460) | |
| | | | **Drug Arrests** | | | | **Other Arrests** | | |
| First-year effect | 2201 | 0.0673 | -0.0004 | -0.0015 | 0.0577 | 0.2219 | -0.0253 | -0.0845 | 0.2280 |
| | | | (0.0145) | (0.0477) | | | (0.0243) | (0.0796) | |
| Average effect | 2201 | 0.0673 | -0.0004 | -0.0015 | 0.0577 | 0.2219 | -0.0253 | -0.0845 | 0.2280 |
| | | | (0.0145) | (0.0477) | | | (0.0243) | (0.0796) | |

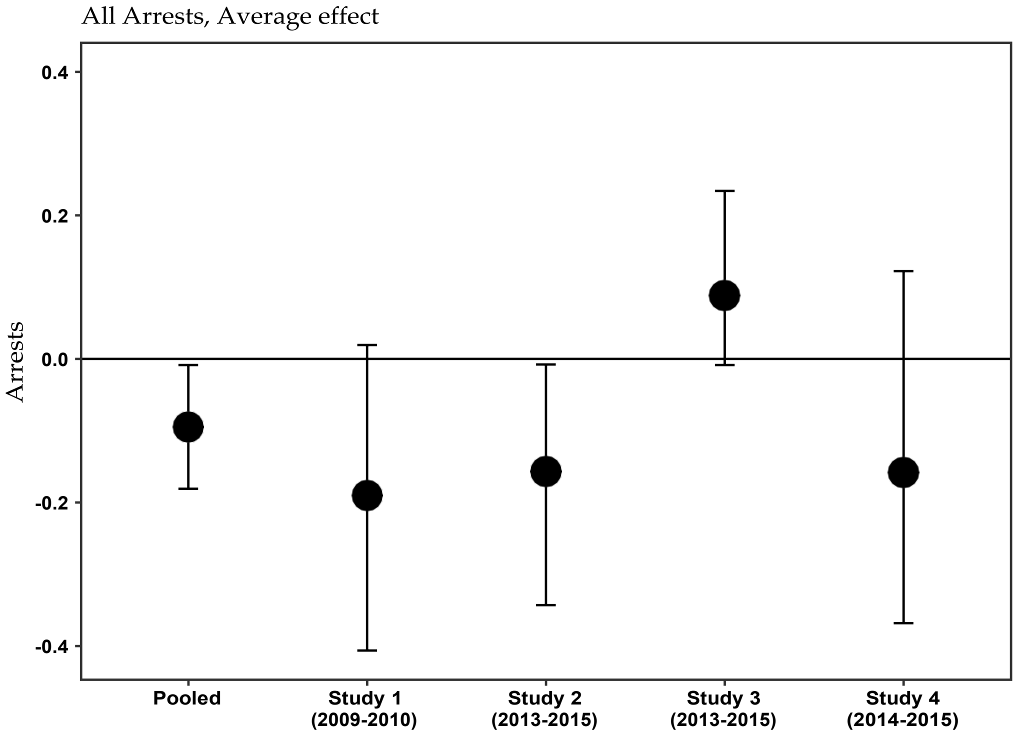
Note: Table presents estimates for the effect of BAM in Study 4. The first-year and average effects measures outcomes from the first year. The second-year effect is omitted because data were not available for the second year of programming in Study 4. The program completion effect is omitted because the BAM curriculum was delivered as a two year program in Study 4 but data were not available to measure outcomes from year 2 when the curriculum would have been completed. Participation for the IV is attending at least one session during the first year. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroscedasticity-robust standard errors clustered on individuals in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

## Figure 1: BAM Impact by Study



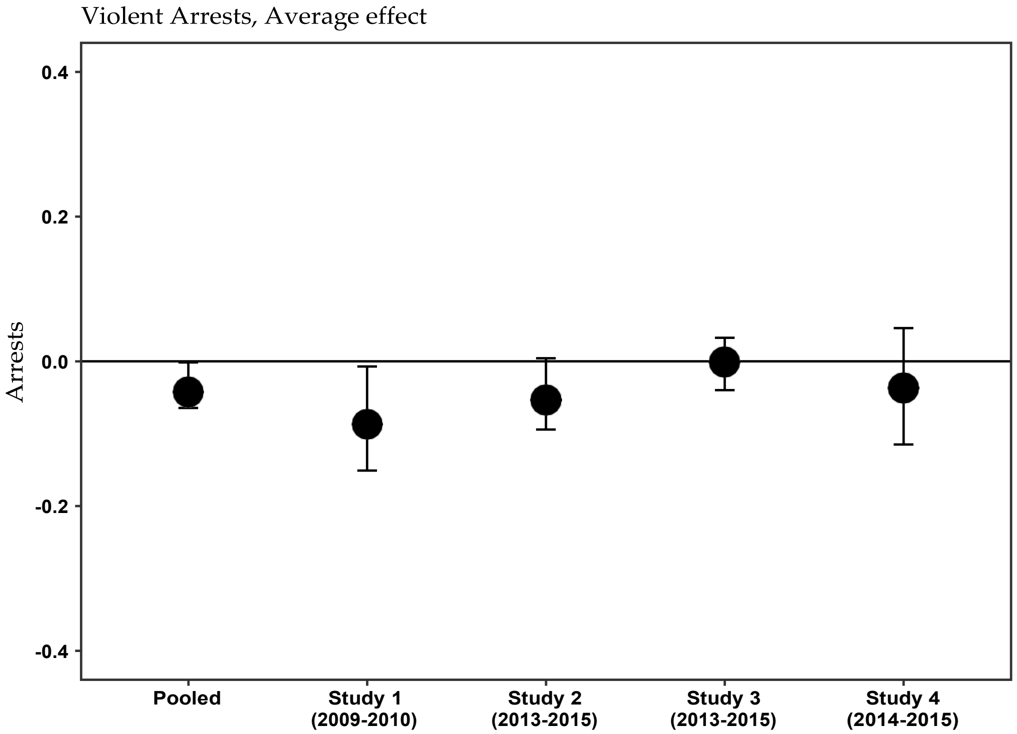
Note: Point estimates are computed as the 50th percentile from a set of 1,000 bootstrap TOT estimates measured in standard deviations and sampled with replacement within randomization block groups. 95% confidence intervals are the 2.5th and 97.5th percentiles of that bootstrap set. The "Pooled" sample includes all participants from each of the four studies. The estimates reflect outcomes from Year 1 of Study 1 and Study 4 and the mean of outcomes from Years 1 and 2 for Study 2 and Study 3. Participation for the IV is measured as attending at least one session during Year 1 for Study 1 and Study 4 and attending at least one session during either year for Study 2 and Study 3. Standard baseline covariates and randomization block fixed effects are included in each model.

## Figure 2: BAM Impact by Study



Note: Point estimates are computed as the 50th percentile from a set of 1,000 bootstrap TOT estimates measured in arrest counts and sampled with replacement within randomization block groups. 95% confidence intervals are the 2.5th and 97.5th percentiles of that bootstrap set. The "Pooled" sample includes all participants from each of the four studies. The estimates reflect outcomes from Year 1 of Study 1 and Study 4 and the mean of outcomes from Years 1 and 2 for Study 2 and Study 3. Participation for the IV is measured as attending at least one session during Year 1 for Study 1 and Study 4 and attending at least one session during either year for Study 2 and Study 3. Standard baseline covariates and randomization block fixed effects are included in each model.

## Figure 3: BAM Impact by Study



Note: Point estimates are computed as the 50th percentile from a set of 1,000 bootstrap TOT estimates measured in arrest counts and sampled with replacement within randomization block groups. 95% confidence intervals are the 2.5th and 97.5th percentiles of that bootstrap set. The “Pooled” sample includes all participants from each of the four studies. The estimates reflect outcomes from Year 1 of Study 1 and Study 4 and the mean of outcomes from Years 1 and 2 for Study 2 and Study 3. Participation for the IV is measured as attending at least one session during Year 1 for Study 1 and Study 4 and attending at least one session during either year for Study 2 and Study 3. Standard baseline covariates and randomization block fixed effects are included in each model.

## Table 11: Heterogeneity in the Effect of BAM Across Studies: Westfall-Young Familywise Error Rate (FWER) Adjustments

**Unadjusted p-Values**

**Panel 1: All outcomes**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| First-year effect | 0.034** | 0.169 | 0.739 | 0.598 | 0.784 | 0.144 |
| Second-year effect | 0.064* | 0.104 | 0.081* | 0.448 | 0.813 | 0.164 |
| Average effect | 0.018** | 0.047** | 0.238 | 0.587 | 0.796 | 0.061* |
| Program completion effect | 0.120 | 0.061* | 0.045** | 0.690 | 0.858 | 0.094* |

**FWER Adjusted p-Values**

**Panel 2: All outcomes**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| First-year effect | 0.416 | 0.712 | 0.862 | 0.862 | 0.862 | 0.712 |
| Second-year effect | 0.566 | 0.677 | 0.623 | 0.862 | 0.862 | 0.712 |
| Average effect | 0.269 | 0.472 | 0.811 | 0.862 | 0.862 | 0.552 |
| Program completion effect | 0.712 | 0.552 | 0.463 | 0.862 | 0.862 | 0.652 |

**Panel 3: All outcomes excluding 'All Arrests'**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| First-year effect | 0.401 | | 0.862 | 0.862 | 0.862 | 0.715 |
| Second-year effect | 0.553 | | 0.611 | 0.862 | 0.862 | 0.722 |
| Average effect | 0.257 | | 0.811 | 0.862 | 0.862 | 0.541 |
| Program completion effect | 0.715 | | 0.447 | 0.862 | 0.862 | 0.638 |

**Panel 4: Average effect outcomes**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| Average effect | 0.114 | 0.183 | 0.558 | 0.798 | 0.798 | 0.217 |

**Panel 5: Average effect outcomes excluding 'All Arrests'**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| Average effect | 0.107 | | 0.558 | 0.798 | 0.798 | 0.217 |

**Panel 6: Program completion effect outcomes**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| Program completion effect | 0.378 | 0.280 | 0.250 | 0.862 | 0.862 | 0.372 |

**Panel 7: Program completion effect outcomes excluding 'All Arrests'**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| Program completion effect | 0.378 | | 0.239 | 0.862 | 0.862 | 0.372 |

**Panel 8: 'School Engagement' outcomes**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| First-year effect | 0.148 | | | | | |
| Second-year effect | 0.168 | | | | | |
| Average effect | 0.099* | | | | | |
| Program completion effect | 0.196 | | | | | |

**Panel 9: 'All Arrests' outcomes**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| First-year effect | | 0.173 | | | | |
| Second-year effect | | 0.173 | | | | |
| Average effect | | 0.114 | | | | |
| Program completion effect | | 0.136 | | | | |

**Panel 10: 'Violent Arrests' outcomes**

| | School Engagement | All Arrests | Violent Arrests | Property Arrests | Drug Arrests | Other Arrests |
|---|---|---|---|---|---|---|
| First-year effect | | | 0.741 | | | |
| Second-year effect | | | 0.171 | | | |
| Average effect | | | 0.308 | | | |
| Program completion effect | | | 0.109 | | | |

Note: Table presents unadjusted p-values from F-tests of treatment effect heterogeneity across study samples and familywise error rate (FWER) adjusted p-values from the same tests  using the Westfall-Young method (1993). Panel 1 presents unadjusted p-values. Panels 2-10 present FWER-adjusted p-values for different outcome families. All models are estimated on sample of pooled BAM studies with standard baseline covariates and randomization block fixed effects. P-values are computed from heteroscedasticity-robust standard errors clustered on students. P values for the first-year effect are computed on outcomes from the first program year of each of the four studies (n=9,307). P-values for the second-year effect are computed on outcomes from the second program year in Study 2 and Study 2x2 and exclude observations from Johnson High School (n=4,174). P-values for the average effect are computed on outcomes from the first program year of Study 1 and Study Expansion an average of outcomes from both program years in Study 2 and Study 2x2 (n=9,307). P-values for the program completion effect are computed on outcomes from the first program year of Study 1 and the second program years of both Study 2 and Study 2x2 (n=6,914).

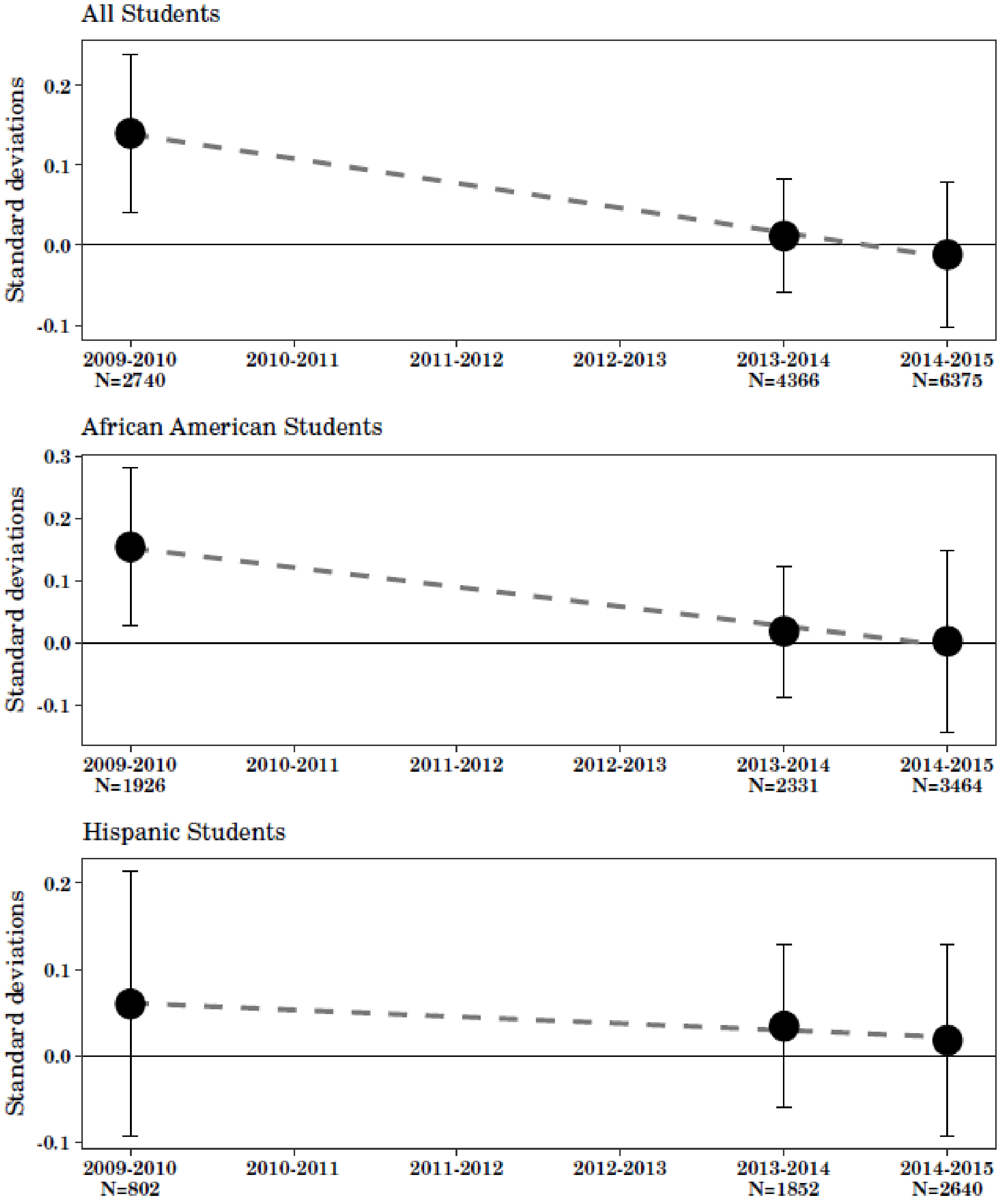
## Table 12: Measuring Moderation and Mediation of the Effect of BAM

| Explanatory Variable | Description |
|---|---|
| **Moderators** | |
| Baseline GPA | Average baseline GPA among all study students at the school |
| Baseline Arrests (count) | Average baseline number of arrests among all study students at the school |
| Baseline Arrests (ever) | Fraction of all study students at the school who had any baseline arrests |
| Baseline Violent Arrests (count) | Average baseline number of arrests for violent crimes among all study students at the school |
| Baseline Suspensions | Average baseline number of out-of-school suspensions for all study students at the school |
| Baseline Annual Attendance | Average number of days present during baseline year for all study students at the school |
| Free and Reduced Lunch | Fraction of all study students at the school who qualify for Free or Reduced price lunch |
| African-American Population | Fraction of all study students at the school who are African-American |
| Hispanic Population | Fraction of all study students at the school who are Hispanic |
| Learning Disability | Fraction of all study students at the school who have a learning disability |
| ELL Status | Fraction of all study students at the school are are English Language Learners |
| Violent Crime Rate I | Violent crime rate per 100k persons in community area where school is located for the first year of study (i.e. 2009 for SY 09-10) |
| Violent Crime Rate II | Violent crime rate per 100k persons in community area where school is located, measured as three-year moving average around the first year of study (i.e. 2009 for SY 09-10) |
| Homicide Rate I | Homicide rate per 100k persons in community area where school is located for the first year of study (i.e. 2009 for SY 09-10) |
| Homicide Rate II | Homicide rate per 100k persons in community area where school is located, measured as three-year moving average around the first year of study (i.e. 2009 for SY 09-10) |
| Public Health & Disadvant age I (alpha=.827) | Z-score index of community area measures of unemployment, HS dropout, teen fertility, and diabetes-related mortality; data are snapshot from 2012 |
| Public Health & Disadvant age II (alpha=.932) | Z-score index of 21 measures of community area-level public health and economic security including and expanding upon the four measures from Index I; data are snapshot from 2012 |
| School Graduation Rate | Schools' graduation rate for the year prior to the study (i.e. SY08-09 for SY 09-10), measured using the pre-2015 method for data consistency |
| School On-Track Status | Schools' 9th grade on-track percentage the first year of study (i.e. SY 08-09 for SY 09-10), measured using the pre-2015 method for data consistency |
| School 5E (Overall) Score | Overall Organization score by school from 5Essentials/My Voice, My School, where 1 is least organized and 5 is most organized |
| School 5E (Ambitious) Score | Ambitious Instruction score by school from 5Essentials/My Voice, My School, where 0 is worst and 100 is best |
| School 5E (Effective) Score | Effective Leaders score by school from 5Essentials/My Voice, My School, where 0 is worst and 100 is best |
| School 5E (Collaborative) Score | Collaborative Teachers score by school from 5Essentials/My Voice, My School, where 0 is worst and 100 is best |
| School 5E (Involved) Score | Involved Families score by school from 5Essentials/My Voice, My School, where 0 is worst and 100 is best |
| School 5E (Supportive) Score | Supportive Environment score by school from 5Essentials/My Voice, My School, where 0 is worst and 100 is best |
| School 5E (Average) Score | Average of the five sub-scores from 5Essentials/My Voice, My School, where 0 is worst and 100 is best |
| **Mediators** | |
| Highest Lesson | The highest lesson out of 30 in BAM curriculum that was taught during the school year; not necessarily a measure of the number of lessons taught throughout the year |
| Attendance | Attendance rate, per counselors session logs, of BAM students across all BAM groups in a given school |
| Attendance (Non-Sports) | Attendance rate, per counselors session logs, of BAM students across all non-sports BAM groups in a given school |

| | |
|---|---|
| Attendance (Sports) | Attendance rate, per counselors session logs, of BAM students across all sports BAM groups in a given school |
| Sessions Held | Average number of sessions held per BAM group in a given school |
| Sessions Held (Non-Sports) | Average number of sessions held per non-sports BAM group in a given school |
| Sessions Held (Sports) | Average number of sessions held per sports-BAM group in a given school |
| Counselor Overall Score | Average counselor score (1-3, best) assigned by Youth Guidance, weighted by counselors' months of service at the school during study-year |
| Clinical Strength Score | Average counselor clinical strength score (1-3, best) assigned by Youth Guidance, weighted by counselors' months of service at the school during study-year |
| Counselor BAM Tenure | Average months since counselors were hired by YG, benchmarked at September of study year and weighted by counselors' months of service at the school during study-year |
| School BAM Tenure | Number of years BAM had been in the school prior to the study-year |
| Counselor Consistency | Measure for whether counselors were replaced (for any reason) during the study-year |
| Counselor Job Experience: Top-Third Disadvantaged Communities | Count of jobs/experiences that were any of the top-third most economically disadvantaged community areas in Chicago, counselor-service weighted |
| Counselor Job Experience: Top-Half Disadvantaged Communities | Count of jobs/experiences that were any of the top-half most economically disadvantaged community areas in Chicago, counselor-service weighted |
| Counselor Job Experience: Community/Service | Count of jobs/experiences that were in community/service related industry (i.e. job training, homeless services, etc.), counselor-service weighted |
| Counselor Job Experience: Schools | Count of jobs/experiences that were in schools or involved cooperating with/navigating school administrations, counselor-service weighted |
| Schools Counselor Job Experience: Kids/Youth | Count of jobs/experiences that involved interactions with kids and youth, counselor-service weighted |
| Counselor Job Experience: Athletics | Count of jobs/experiences that were in athletics, counselor-service weighted |
| Counselor Job Experience: Religious | Count of jobs/experiences that were in religious organizations, counselor-service weighted |
| Counselor Job Experience: Counseling | Count of jobs/experiences that involved direct or group counseling (position had to include actual performance of counseling, rather than being tangential to or in service of counseling), counselor-service weighted |
| Counselor Degree Level | Counselors' highest degree level achieved, counselor-service weighted |
| Counselor Psych/Clinical Training | Degree in counseling or related field, by level of degree, counselor-service weighted |
| Counselor Age (est.) | Estimated age of counselor based on HS/College degree award years, counselor-service weighted |

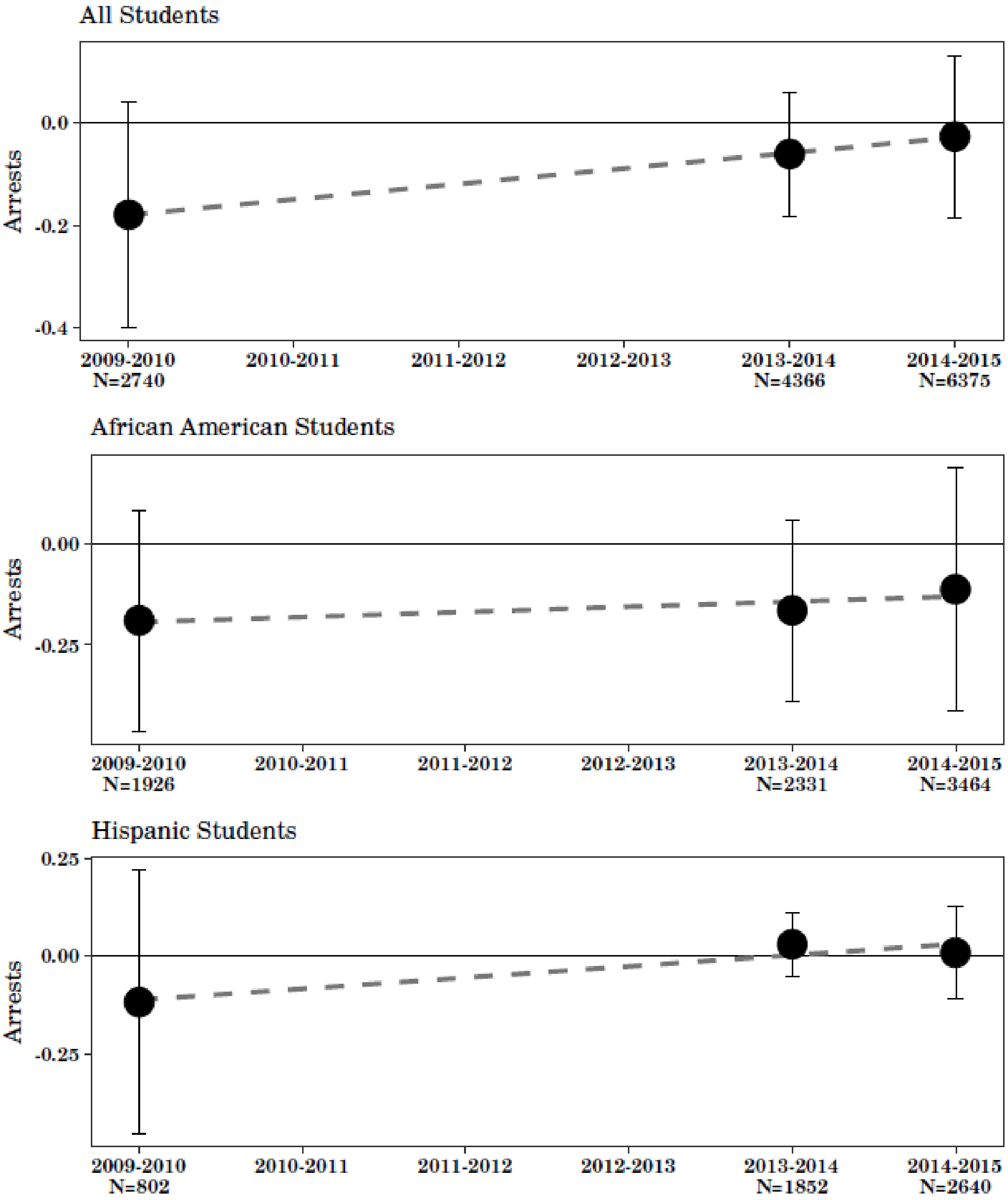
Note: The variables above are each measured at the school-year level. Some are constant within school-years (e.g. Homicide Rate) while others are averaged across each student in a given school-year (e.g. Baseline GPA).

**Figure 4: BAM Impact on School Engagement, by student race**



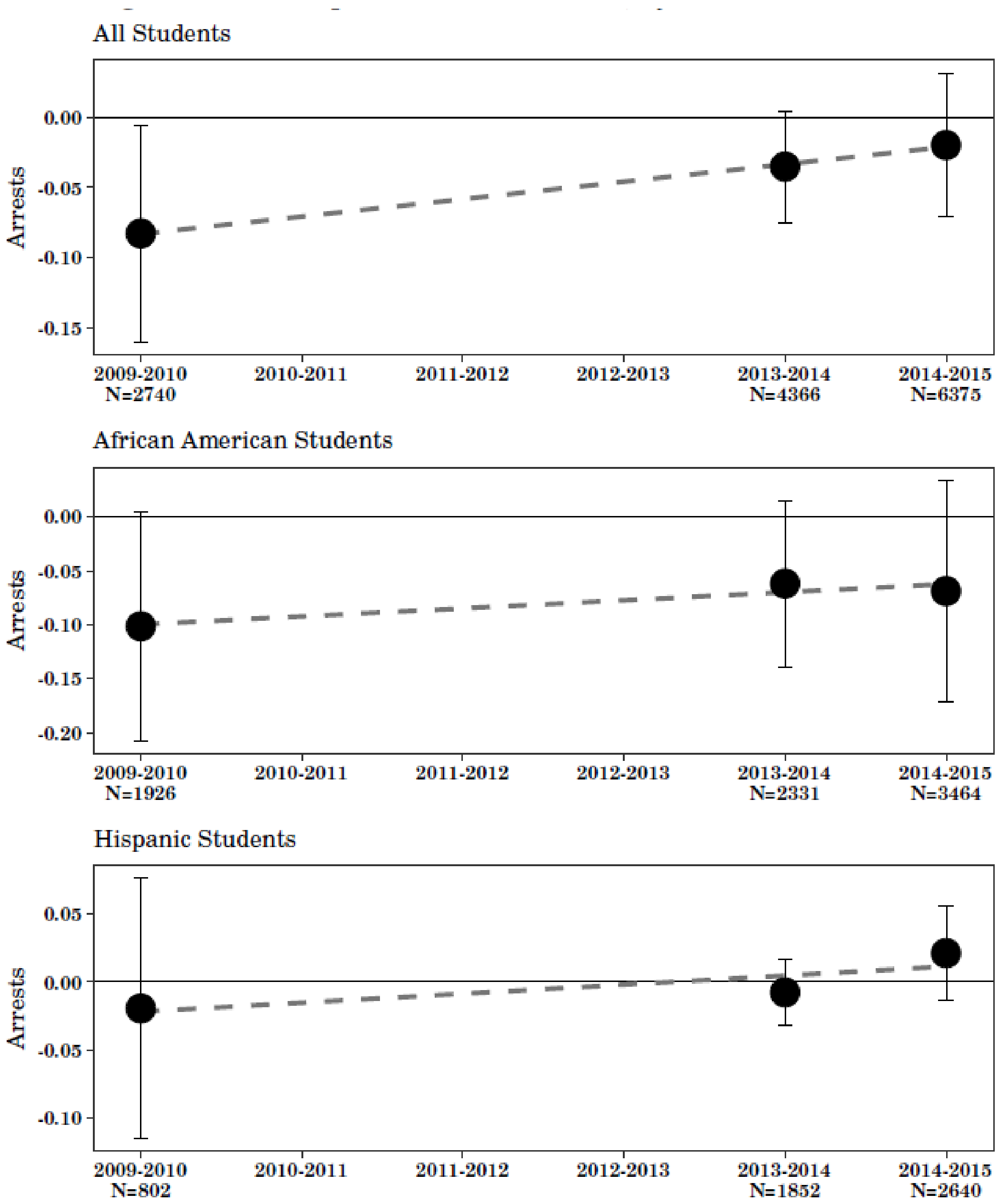
Note: TOT point estimates for the effect of BAM in a given study year are displayed as a function of students' race. School engagement outcomes are measured in standard deviations. Participation for the IV is measured as attending at least one session during the year. The estimates for 2009-10 include participants from Study 1 only; the estimates for 2013-14 include participants from Study 2 and Study 3; the estimates for 2014-15 include participants from Study 2, Study 3, and Study 4. A small number of students who are neither black nor Hispanic are included in the first panel ("All Students"). Dashed lines reflect the linear best fit. Standard baseline covariates and randomization block fixed effects are included in each model and standard errors are clustered at the student-level.

**Figure 5: BAM Impact on All Arrests, by student race**



Note: TOT point estimates for the effect of BAM in a given study year are displayed as a function of students' race. Outcomes are measured in arrest counts. Participation for the IV is measured as attending at least one session during the year. The estimates for 2009-10 include participants from Study 1 only; the estimates for 2013-14 include participants from Study 2 and Study 3; the estimates for 2014-15 include participants from Study 2, Study 3, and Study 4. A small number of students who are neither black nor Hispanic are included in the first panel ("All Students"). Dashed lines reflect the linear best fit. Standard baseline covariates and randomization block fixed effects are included in each model and standard errors are clustered at the student-level.

## Figure 6: BAM Impact on Violent Arrests, by student race



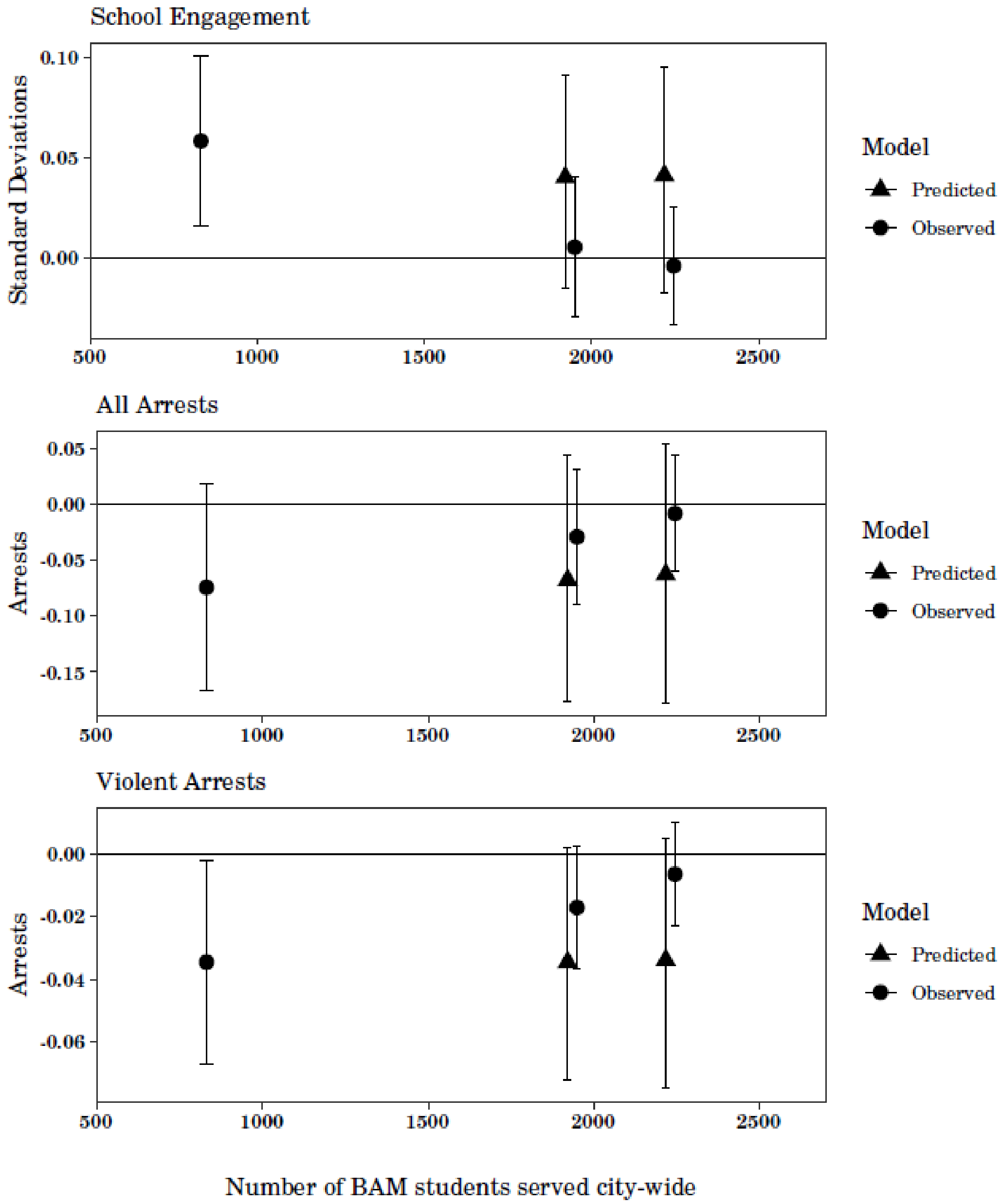
Note: TOT point estimates for the effect of BAM in a given study year are displayed as a function of students' race. Outcomes are measured in arrest counts. Participation for the IV is measured as attending at least one session during the year. The estimates for 2009-10 include participants from Study 1 only; the estimates for 2013-14 include participants from Study 2 and Study 3; the estimates for 2014-15 include participants from Study 2, Study 3, and Study 4. A small number of students who are neither black nor Hispanic are included in the first panel ("All Students"). Dashed lines reflect the linear best fit. Standard baseline covariates and randomization block fixed effects are included in each model and standard errors are clustered at the student-level.

### Table 13: Effect of BAM – Observed and Predicted by Program Year

| Outcome | Observed | Predicted | | |
|---|---|---|---|---|
| | ITT | Using moderators | Using mediators | Using all features |
| **Effect of BAM in 2014** | | | | |
| School Engagement | 0.0055 | 0.0400 | 0.0471 | 0.0403 |
| All Arrests | -0.0292 | -0.0687 | -0.0439 | -0.0677 |
| Violent Arrests | -0.0171 | -0.0362 | -0.0455 | -0.0346 |
| **Effect of BAM in 2015** | | | | |
| School Engagement | -0.0038 | 0.0408 | 0.0498 | 0.0413 |
| All Arrests | -0.0083 | -0.0622 | -0.059 | -0.0627 |
| Violent Arrests | -0.0063 | -0.0345 | -0.0547 | -0.0339 |

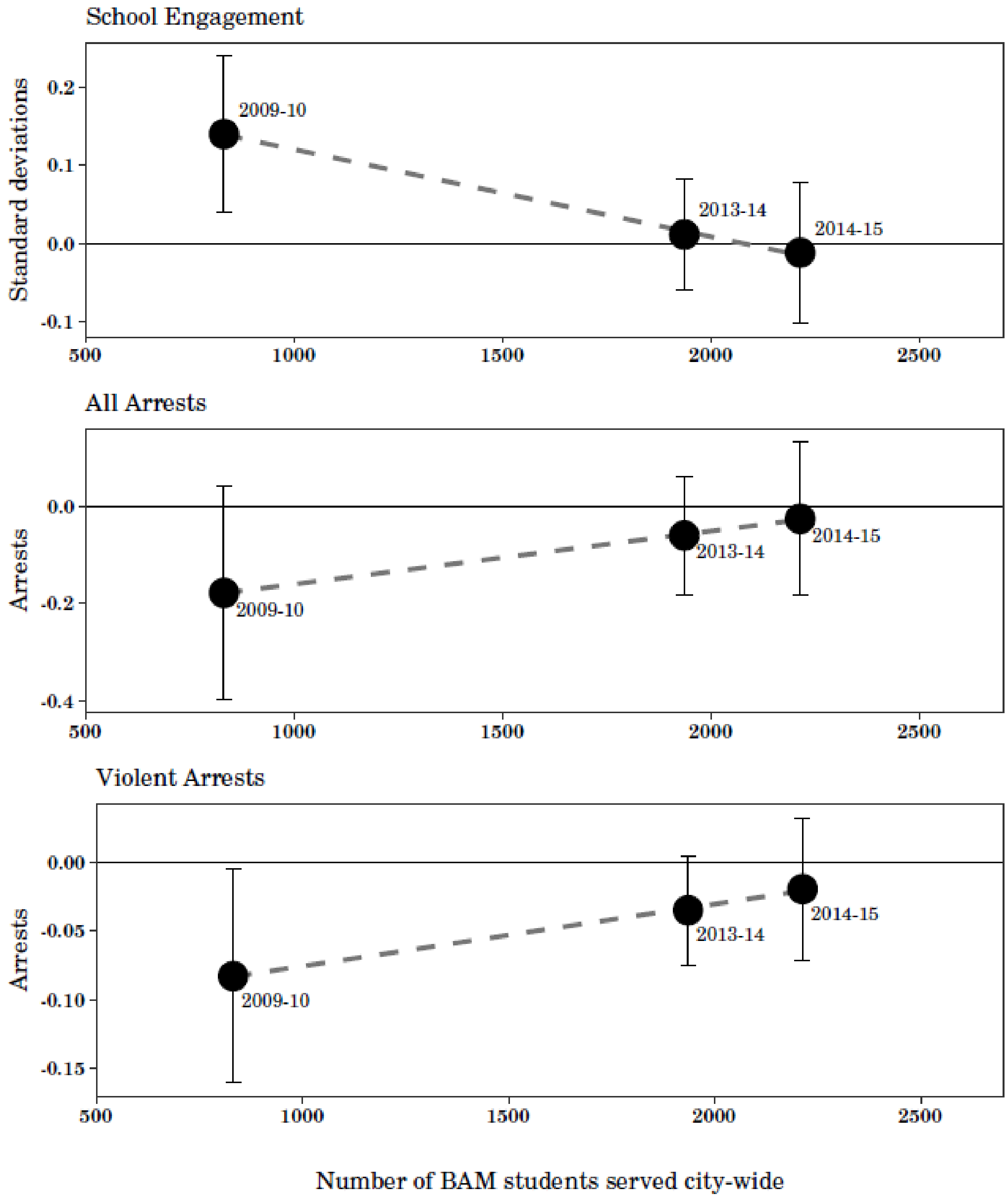
Note: Table presents observed and predicted point estimates for the ITT effect of BAM in 2013-14 and 2014-15. Observed estimates include standard baseline covariates and randomization block group fixed effects as well as standard errors clustered at the student-level. Predicted estimates are measured as the mean of 30,000 bootstrap replications training a generalized random forest algorithm (grf) using random sampling with replacement. We train the grf using only data from the 2009-10 program year -- Study 1. The trained model returns predicted estimates for the effect of BAM at the individual-level based on a set of features describing students' personal, school, and neigborhood characteristics (moderators) as well as characteristics describing the implementation of BAM at their school during a given program year (mediators). We make different predictions for the effect of BAM by training the model using moderators only, mediators only, and the full set of features. Taking an average of each student's predicted treatment effect in a given program year, we can derive and average predicted treatment effect for the program year comparable to our observed ITT estimate for that same year.

## Figure 7: Effect of BAM, observed and predicted estimates



Note: See note to Table 13 for further description of the predictive model. Point estimates are presented as a function of the number of BAM students served city-wide in a given year, including non-study students. Circular points reflect observed ITT outcomes for study students. Triangular points reflect predicted outcomes among students in the 2013-14 and 2014-15 study years based on a model trained using observed outcomes from only the 2009-10 study year. Predicted outcomes were computed as the 50th percentile from a set of 30,000 bootstrap estimates, sampled with replacement. 95% confidence intervals reflect observed standard errors for the ITT and the 2.5th and 97.5th percentiles of the bootstrap set for the predicted estimates. School engagement outcomes are measured in standard deviations while arrest outcomes are measured in arrest counts. Dashed lines reflect the linear best fit. ITT models include standard baseline covariates and randomization block fixed effects as well as standard errors clustered at the student-level.

**Figure 8: BAM Impact by Scale of BAM Delivery City-Wide**



Note: TOT point estimates are displayed as a function of the number of BAM students served city-wide in a given year. Point estimates measure outcomes among study participants while the number served city-wide includes all BAM participants, including those outside of the study sample. School engagement outcomes are measured in standard deviations while arrest outcomes are measured in arrest counts. Participation for the IV is measured as attending at least one session during the year. The estimates for 2009-10 include participants from Study 1 only; the estimates for 2013-14 include participants from Study 2 and Study 3; the estimates for 2014-15 include participants from Study 2, Study 3, and Study 4. Dashed lines reflect the linear best fit. Standard baseline covariates and randomization block fixed effects are included in each model and standard errors are clustered at the student-level.

## IX. REFERENCES

Anderson, E. (1999). *Code of the Streets*. New York, NY: Norton.

Anderson, M.L. (2008). Multiple inference and gender differences in the effects of early intervention: reevaluation of the Abecedarian Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association,* 103(2008), 1481-95.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434), 444-455.

Beck, J. (2011). *Cognitive Therapy: Basics and Beyond*. New York, NY: The Guilford Press.

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

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

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation review*, 8(2), 225-246.

Borghans, L., Duckworth, A.L., Heckman, J.J., & Ter Weel, B. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources,* 43(4), 972-1059.

Bowles, S., Gintis, H., & Osborne, M. (2001). The determinants of earnings: A behavioral approach. *Journal of Economic Literature*, 39(4), 1137-1176.

Carneiro, P. & Heckman, J. (2003). Human Capital Policy. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 9495.

Chalfin, A. & McCrary, J. (2017). Are U.S. Cities Underpoliced?: Theory and Evidence. *The Review of Economics and Statistics*.

Cohen M., Rust, R., Steen, S., & Tidd, S. (2004). Willingness-to-Pay for Crime Control Programs. *Criminology*, 42(1), 89-110.

Cook, P., Dodge, K., Farkas, G., Fryer, R., Guryan, J., Ludwig, J., Mayer, S. (2015). Not Too Late: Improving Academic Outcomes for Disadvantaged Youth. *Northwestern University Institute for Policy Research Working Paper Series.*

Cunha, F. & Heckman, J.J. (2007). The technology of skill formation. *American Economic Review*, 97(2), 31-47.

Cutler, D.M. & Lleras-Muney, A. (2008). Chapter 2: Education and Health: Evaluating

Theories and Evidence," in *Making Americans Healthier: Social and Economic Policy as Health Policy*. New York: Russell Sage Foundation.

Deming, D. J. (2011). Better Schools, Less Crime?. *The Quarterly Journal of Economics,* 126 (4), 2063–2115.

Durlak, J. A., Dymnicki, A. B., Taylor, R. D., Weissberg, R. P., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, 82(1), 405-32.

Evans, W.N., & Owens, E.G. (2007). COPS and Crime. *Journal of Public Economics,* 91(1-2), 181-201.

Goldin, C. & Katz, L.F. (2008). *The Race between Education and Technology.* Cambridge, MA: Harvard University Press.

Hammond, W. R. & Yung, B. (1991). Preventing Violence in At-Risk African-American Youth. *Journal of Health Care for the Poor and Underserved*, 2(3), 359-373.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4), 451-464.

Heckman, J. J., & LaFontaine, P. A. (2010). The American High School Graduation Rate: Trends and Levels. *Review of Economics and Statistics,* 92(2), 244-262.

Heckman, J. J., & Rubinstein, Y. (2001). The Importance of Noncognitive Skills: Lessons from the GED Testing Program. *American Economic Review*, 91(2), 145-149.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics,* 24(3), 411-482.

Heller, S. B., Pollack, H. A., Ander, R., & Ludwig, J. (2013). Preventing youth violence and dropout: A randomized field experiment. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 19014.

Heller, S.B., Shah, A.K., Guryan, J., Ludwig, J., Mullainathan, S. & Pollack, H. (2017). Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago. *Quarterly Journal of Economics,* 132(1), 1-54.

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).

Jones, S. M., Brown, J. L., & Aber, J. L. (2011). Two-year impacts of a universal school-based social-emotional and literacy intervention: An experiment in translational developmental research. *Child Development*, 82(2), 533-554.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Katz, L.F., Kling, J.R., and Liebman, J.B. (2001). Moving to opportunity in Boston: Early results of a randomized mobility experiment. *Quarterly Journal of Economics*, 116(2), 607-654.

Kling, J.R., Liebman, J.B., & Katz, L. (2007). Experimental Analysis of Neighborhood Effects. *Econometrica*, 75(1), 83-119.

Kling, J. R., Ludwig, J. & Katz, L.F. (2005). Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment. *Quarterly Journal of Economics,* 120(1), 87-130.

Little, R.J. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd Edition.* Wiley Series in Probability and Statistics.

Lleras-Muney, A. (2005). The Relationship between Education and Adult Mortality in the United States. *Review of Economic Studies,* 72(1), 189-221.

Lochner, L. & Moretti, E. (2004). The Effect Of Education On Crime: Evidence From Prison Inmates, Arrests, And Self-Reports. *American Economic Review*, 94(1), 155-189.

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., Sanbonmatsu, L. (2012). Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults. *Science*, 337(6101) 1505-1510.

Miller, T., Cohen, A., & Wiersema, B. (1996). *Victim Costs and Consequences: A New Look.* U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, 108(7), 2693–2698.

Monahan, K.C., Steinberg, L., Cauffman, E., & Mulvey, E.P. (2009). Trajectories of antisocial behavior and psychosocial maturity from adolescence to young adulthood. *Developmental psychology*, 45(6), 1654.

Murnane, R. J. (2013). U.S. high school graduation rates: Patterns and explanations. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 18701.

Papachristos, A.V. (2009). Murder by Structure: Dominance Relations and the Social Structure of Gang Homicide. *American Journal of Sociology*, 115(1): 74-128.

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

Reardon, S. F. (2011). "The widening academic achievement gap between the rich and the poor: New evidence and possible explanations." In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, Eds. Greg J. Duncan and Richard J. Murnane. New York: Russell Sage Foundation Press. pp. 91-116.

Schochet, P.Z., Burghardt, J., McConnell, S. (2008). Does Job Corps work? Impact findings from the National Job Corps Study. *American Economic Review*, 98(5), 1864-86.

Thaler, R.H. & Sunstein, C.R. (2008). *Nudge: Improving Decisions about Health, Wealth and Happiness*. New York, NY: Penguin Books.

Tough, P. (2012). *How Children Succeed: Grit, Curiosity, and the Hidden Power of Character*. New York: Houghton Mifflin Harcourt.

Weiner, D.A., Lutz, B.F. & Ludwig, J. (2009). The Effects of School Desegregation on Crime. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 15380.

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