Institute for Policy Research
Northwestern University
Working Paper Series

WP-14-03

The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth from a Randomized Experiment in Chicago

Philip J. Cook and Kenneth Dodge
Duke University

George Farkas
University of California, Irvine

Roland G. Fryer, Jr.
Harvard University

Jonathan Guryan
Assistant Professor of Human Development and Social Policy
Faculty Fellow, Institute for Policy Research
Northwestern University

Jens Ludwig, Susan Mayer, and Harold Pollack
University of Chicago

Laurence Steinberg
Temple University

Version: January 2014

DRAFT
Please do not quote or distribute without permission.

Abstract

There is growing concern that improving the academic skills of disadvantaged youth is too difficult and costly, so policymakers should instead focus either on vocationally oriented instruction for teens or else on early childhood education. Yet this conclusion may be premature given that so few previous interventions have targeted a potential fundamental barrier to school success: "mismatch" between what schools deliver and the needs of disadvantaged youth who have fallen behind in their academic or non-academic development. This paper reports on a randomized controlled trial of a two-pronged intervention that provides disadvantaged youth with non-academic supports that try to teach youth social-cognitive skills based on the principles of cognitive behavioral therapy (CBT), and intensive individualized academic remediation. The study sample consists of 106 male 9th and 10th graders in a public high school on the south side of Chicago, of whom 95% are black and 99% are free or reduced price lunch eligible. Participation increased math test scores by 0.65 of a control group standard deviation(SD) and 0.48 SD in the national distribution, increased math grades by 0.67 SD, and seems to have increased expected graduation rates by 14 percentage points (46%). While some questions remain about the intervention, given these effects and a cost per participant of around $4,400 (with a range of $3,000 to $6,000), this intervention seems to yield larger gains in adolescent outcomes per dollar spent than many other intervention strategies.

I. INTRODUCTION

By age 13 the gap in achievement test scores between African-American and white children, as measured in the National Assessment of Educational Progress (NAEP), equals 0.62 standard deviations (SD) in reading and 0.80SD in math.[1] Disparities in test scores along income lines are even larger and are growing over time (Reardon, 2011). Inequality in academic achievement is an important contributor to other forms of inequality, for example with respect to schooling attainment, income, health, and crime involvement.

While there is widespread agreement about the importance of this problem, there remains great uncertainty about the best way to solve it. There are remarkably few success stories of efforts to improve academic outcomes of disadvantaged youth,[2] which has led to growing concerns about the value of such efforts. For example Cullen, Levitt, Robertson and Sadoff (2013) argue that rather than focus on college-bound academics for disadvantaged teens, secondary schools should focus on technical or vocational education. Carniero and Heckman (2003, p. 90) argue for a focus on younger children: "The return to [human capital] investment in the young is apparently quite high; the return to investment in the old and less able is quite low."

Yet the conclusion that adolescence is too late to improve the academic outcomes of disadvantaged children may be premature, given the possibility that previous interventions may have misdiagnosed the key barriers to success for this population and so have been aiming at the wrong target. The U.S. currently spends around $590 billion each year on public K-12 schooling.[3] After the first few grades, the explicit focus of most public school instruction is on the development of academic skills. Most education reform efforts focus on improving the quality with which grade-level material is taught, or the incentives students have to learn it.

We hypothesize that there are important mismatches between what many students (especially those from disadvantaged backgrounds) need, and what many current education policies try (or are able) to provide.

---

[1] The exact magnitude of the black-white gap depends on the study sample examined, the age at which the gap is measured, the achievement assessment that is used, and the academic subject being examined; most studies report the gap among adolescents to be in the range from 0.5 to 0.9 standard deviations, with gaps that tend to be larger for math than reading (Jencks & Phillips, 1998; Clotfelter, Ladd & Vigdor, 2009; Fryer, 2010; Reardon, 2011).

[2] Most academic interventions for disadvantaged adolescents tend to focus on measures like schooling attainment or high school graduation as the outcome of primary interest. The U.S. Department of Education's What Works Clearinghouse gives no dropout prevention program its top rating for strong effects, while the Coalition for Evidence-Based Policy does not list a single program for addressing high-school graduation rates among its "Top Tier" programs.

[3] http://www.census.gov/compendia/statab/2012/tables/12s0261.pdf

Growing up in distressed, dangerous urban areas may affect the development of "non-academic" factors like social information processing styles or other features of judgment and decision-making that affect how students engage with school throughout their K-12 careers. These students may benefit from additional help with decision making at ages where schools no longer explicitly focus on this area.

On the academic side we know that the variance in achievement among all students increases as they progress in school (Cascio and Staiger, 2012), a problem that may be exacerbated in urban areas where severe disadvantage among many students affects the rate at which they learn academic material, which then leads them to fall behind grade level, which makes it more difficult to keep up with subsequent grade-level instruction, which leads them to fall yet further behind. Assessments of some of the most disadvantaged young people in Chicago – those who wind up in the Cook County Jail – find they can be up to seven years behind grade level in reading and up to 10 years behind in math (Keeley, 2011). The need for options to intensively help those who have fallen behind – to provide a real safety net – is a key systemic challenge for urban school districts serving large numbers of disadvantaged students.

In this paper we report on the results of a randomized controlled trial (RCT) of a two-pronged intervention that tries to provide both academic and non-academic remediation for disadvantaged youth who are falling behind and at great risk for slipping through the cracks of the current school system and dropping out. The non-academic intervention, developed by Youth Guidance and called Becoming a Man (BAM), includes social-cognitive skills training based on cognitive behavioral therapy (CBT) principles. A previous RCT by our team found the intervention reduced rates of violence involvement (by 44%) and increased schooling engagement, but did not have detectable effects on test scores when delivered on its own (Heller et al., 2013).

The academic intervention is intensive, individualized two-on-one math tutoring provided for one hour per day each and every day, based on the model developed by Match Education. For decades education researchers have thought that small-group tutoring generates "the best learning conditions we can devise," and have struggled to solve the key challenge that small-group tutoring by regular teachers is "too costly for most societies to bear on a large scale" (Bloom, 1984, p. 4). Match solves this problem by recognizing that small-group tutoring simplifies the teaching task in many ways, for example by eliminating the need for specialized

training in classroom management, and so greatly expands the set of people capable of being successful instructors. Match hires well-educated committed people who usually do not have formal teacher training, but are willing to work for a year in this job for a modest stipend as a public service (similar to programs like Teach for America). Fryer's (2011) non-experimental study in the Houston Public Schools found gains in math scores in the grades exposed to Match tutoring (6^{th} and 9^{th}) on the order of 0.48SD and 0.74SD, respectively, although that promising intervention has not yet been subject to an RCT.

The data from our Chicago experiment suggests that program participation (the effects of treatment on the treated, or TOT) increased math achievement test scores by 0.65 of the control group's standard deviation (equal to 0.48SD within the national distribution), which equals a change in rank within the national test-score distribution of 15 percentile points. These gains were measured on a broad test of math achievement (ACT Inc.'s EXPLORE and PLAN tests). Participation also improved math grades by 0.67SD, and had sizable (but sometimes not quite statistically significant) effects in reducing absences by one-quarter and F's in math and non-math classes by two-thirds. Participation also improved a Chicago Public Schools (CPS) indicator for being "on track" for graduation (Allensworth and Easton, 2005) by 46%, which translates into a gain in expected high school graduation rates of about 14 percentage points.

These results are striking partly because they come from working with a target population of the sort for which many have thought improving academic outcomes was infeasible – 106 male youth enrolled in 9^{th} and 10^{th} grade in academic year 2012-13 in a public high school on the south side of Chicago. Of the youth in our study sample, 99% were eligible for free or reduced price lunch and 95% are black, with average baseline reading and math scores that fell at the 26^{th} and 22^{nd} percentiles of the national distribution, respectively. What is also striking about these results is that programming did not start until the middle of November, so the programming duration was only about three-quarters of an academic year.

There are several important questions that remain about our results, including whether or how these results will persist over time, and the relative effectiveness of the two components of the intervention. The fact that there appears to be some crossover or spillover across the academic and non-academic components of our intervention, together with our small sample size, makes it complicated to cleanly distinguish the effects of the

different components of the intervention bundle. Yet our benchmark estimate for the combined cost of the "treatment bundle" is on the order of $4,400, with a defensible range of $3,000 to $6,000 per student. While such a small study cannot be the definitive word on how to reduce disparities in academic outcomes within the U.S., if these pilot results could be achieved at large scale the gains in adolescent outcomes per dollar spent would be larger than many other strategies that have been tried, including in the early childhood area.

The remainder of this paper is organized as follows: The second section discusses the theory behind the interventions we deliver in this RCT. The third section describes the interventions. Our data sources are described in section four; our analytic approach is outlined in section five; our main findings are reported in section six; and the limitations and implications of these results, including how the gains per dollar spent from this intervention compare with other educational interventions, are discussed in section seven.

II. THEORY

Our study is motivated by the hypothesis that there is a "mismatch" between the sorts of supports that disadvantaged youth need to succeed in school, and what most previous education or social policy interventions have provided. That mismatch, we believe, provides an explanation for why so few previous interventions have been successful – which runs counter to the alternative hypothesis that adolescence is already too late to intervene and substantially and cost-effectively improve academic outcomes.

A. Non-academic barriers

To understand the sorts of supports disadvantaged youth in our study site of Chicago might benefit from, it is first useful to understand the context in which these youth are growing up and attending school. The CPS system is one of the nation's largest urban school districts, with over 23,000 teachers serving over 400,000 students in 681 schools, including 106 high schools. As in many urban districts, students are disproportionately from disadvantaged family backgrounds. Fully 87% of CPS students are eligible for free or reduced price lunch; 42% are African-American, 44% are Latino, 9% are white, and 3% are Asian/Pacific-Islander.

The specific high school in which we carried out the present study's RCT is located in a very racially and economically segregated neighborhood on the south side of Chicago. Mirroring the socio-demographics of the surrounding community, nearly all students in the study high school are African-American and are eligible

for free or reduced price lunch. The school is relatively small compared to many other big-city high schools, with fewer than 1,000 total students enrolled across grades 9-12. The community in which the school is located is also among the most dangerous in Chicago. The homicide rate in the community in recent years has ranged from around 35 to 55 per 100,000. By way of comparison, the homicide rate in Chicago as a whole in recent years has been around 15 or 18 per 100,000, and in the U.S. as a whole is on the order of 5 or 6 per 100,000.

One key reason many people have become skeptical about the efficacy of academic programming for disadvantaged youth is the concern that the deleterious effects of poverty may already be too entrenched by adolescence, so that broader policy responses are required that also address the non-academic barriers to school success that children from low-income backgrounds face (see for example Ladd, 2012). The powerful role of family background in explaining how students fare in school has been a major education-policy concern since at least the time of the landmark Coleman Report of 1966 (Coleman et al., 1966; see also Jencks and Mayer, 1998, and Duncan and Murnane, 2011).

In principle one way to address the harmful effects of poverty and disadvantage on children's schooling outcomes is to directly reduce poverty and disadvantage, although this has proven challenging to achieve in practice. The official poverty rate in the U.S. has not changed much since the late 1960s (see Figure 4 in DeNavas-Walt, Proctor and Smith, 2011). While focusing on consumption rather than cash income suggests a somewhat more promising long-run trend in poverty rates (Meyer and Sullivan, 2013), the sorts of transfer policies that may help further reduce poverty in America are costly.[4]

An alternative approach is to try to address the mechanisms that mediate the deleterious effects of poverty and disadvantage on schooling outcomes. A large body of correlational research shows that schooling and other key life outcomes are correlated with what economists have come to call "non-cognitive skills" (or what psychologists like Dodge et al. (1986) call "social-cognitive skills") such as self-regulation, social

[4] Many social scientists have long thought that community-level disadvantage is also an important determinant of children's learning outcomes. While the amount of residential segregation by race in the U.S. has been declining over time since 1970 (Glaeser and Vigdor, 2012), the amount of income segregation has been increasing; see Kneebone, Nadeau and Berube (2011), Watson (2009) and Reardon and Bischoff (2011).

information processing, conflict resolution, "grit" and future orientation.[5] A focus on non-academic barriers to success is thought by many "no-excuses" charter schools to be one of the keys to their own perceived success; as the dean of students at one school said, "At KIPP, we've always said that character is just as important as academics" (Tough, 2012, p. 86).

Yet most public school systems at present do not devote much time to explicitly addressing non-academic factors, at least after the earliest grades of elementary school. Previous research does find that spending more time in high school seems to increase non-academic factors related to decision-making, trust, and risky behavior (Oreopoulos and Salvanes, 2011). But there may be high returns to devoting more explicit attention and effort to addressing these non-academic barriers to school success during adolescence. The quality of the existing empirical evidence on the value of explicitly addressing non-academic factors or skills among disadvantaged youth is currently not ideal.[6] This remains an important open question.

B. Academic barriers

Given the high levels of disadvantage that so many children in Chicago and other American cities face, it is perhaps not surprising that many struggle to keep up in school – although there is a substantial amount of variation in the degree to which children fall behind. In general, education data show that the variance in student achievement increases as children progress in school (Cascio and Staiger, 2012). The result is great variability in academic levels and needs by middle or high school, which are quite pronounced in urban school districts like CPS. In the 2011 NAEP, fully 40% of 8th graders in Chicago were below basic level in math, 40% were at basic level, 17% were at proficient level and 3% were advanced.[7] Keeley (2011) found that among those Chicago youth at highest risk for school failure and crime (those arrested and sent to the Cook County Jail),

[5] See also Borghans, et al. 2007; Bowles, Gintis & Osborne 2001; Cunha & Heckman 2007; Dodge 2003; Dodge, et al., 1986; Heckman & Rubinstein 2001; Heckman, Stixrud & Urzua 2006; Moffitt, et al. 2011; Monahan, et al. 2009.

[6] Social emotional learning (SEL) programs have shown mixed results with youth in school settings, which makes it difficult to pin down reasonable intervention strategies or expectations of longer-term results. For example the Positive Adolescent Choices Training (PACT) intervention helps African-American youth better interact with each other and finds that there is a reduction in school suspensions as a result (Hammond & Yung, 1991). On the other hand, consider the RCT of the 4Rs program ("reading, writing, respect and resolution"), which provided a 21- to 35-lesson literacy-based SEL curriculum and 25 hours of teacher training and ongoing coaching. Jones et al. (2011) report 50 different impact estimates (intercepts and slopes for main effects, as well as interactions with baseline covariates) out of which just four were significant at 95% (there is about a one in seven chance we'd see that just by chance if these were all independent tests). Meta-analyses like Durlak, et al. (2011) are more positive about SEL programs overall but more than half the studies included there are not RCTs, and results from only RCTs are not reported separately. For a more detailed discussion see Heller, et al. (2013).

[7] http://nationsreportcard.gov/math_2011/math_2011_tudareport/

some had academic skills at grade level. But on average these youth were two years behind grade level in reading, with some up to seven years behind, and four years behind grade level in math, with some having math skills fully 10 years below grade level.

This substantial variation in academic level among disadvantaged youth in Chicago (and other cities) may create a "mismatch" between what many students need and what is delivered in regular classroom settings.[8] Some empirical support for this "academic mismatch" hypothesis comes from Duflo et al.'s (2011) study in Kenya that randomly assigned schools to continue status quo operations or else to group students into classrooms based on academic achievement level. Learning was higher in "tracked" schools for students *both* in the top *and* bottom halves of the achievement distribution. This experiment suggests that for initially low-performing students the benefits in tracked schools from reduced academic mismatch (better-targeted instruction) are not only important, but also large enough to outweigh any adverse peer effects from having lower-achieving classmates in tracked schools.

Of course tracking is not necessarily the only – or necessarily the best – solution to the problem of academic mismatch. An alternative approach would be to bring students at the bottom of the achievement distribution up closer to grade level so that it would be easier to deliver instruction matched to more students' skills within a classroom setting. Unfortunately most urban public school systems are currently not well equipped to individualize academic instruction to the extent necessary to bring students who are already farthest behind up to grade level.

Tracking involves both reducing the mismatch between the skill level of students and the material being taught, and grouping students into fixed groups according to skills assessed at some point in time. The latter has the drawback of being inflexible, and of potentially reducing the possibility of upward mobility among students later in their academic career. It is possible to provide the former without the latter by individualizing instruction. Some evidence for the potential value of individualized, intensive remediation comes from the RCT

[8] Previous research suggests there can be mismatches between the developmental needs of youth and their social environments, also called "stage-environment fit" (see Hunt, 1975, Eccles et al. 1993). The same sort of mismatch may occur for youth's academic needs as well. For example Engel, Claessens and Finch (2012) find that there is mismatch in math instruction among young children in the opposite direction to what we study here – namely, that many kindergarten classrooms teach math content that is too easy, which children already know.

carried out by Banerjee et al. (2007), which found that assigning third and fourth graders who are far behind to receive instruction in remedial academic skills for two hours per day in a classroom of 15-20 students increased test scores by around 0.6 SD. Interestingly, given the growing focus in the U.S. on the importance of teacher "quality," the instructors for these remedial classes were women from the local community who were trained for just a short period of time and paid only $10-15 per month. The effects of a computer-assisted program that also helped individualize instruction was found to increase test scores by up to 0.47 SD after the second year of intervention, although impacts from either strategy were short-lived.

Some non-experimental but highly suggestive support for the value of individualized remedial instruction in the U.S. context comes from Fryer's (2011) study of "no excuses" charter school reforms in the Houston public schools. Fryer identified five promising features of "no excuses" charter schools, four of which were provided to all students in a selected set of Houston Public Schools in AY2010-11, while 6th and 9th graders in these schools also received two-on-one math tutoring for one hour per day every day delivered by Match Education of Boston. Fryer estimated the effects on students using both a difference-in-difference approach, comparing pre-post trends in treatment vs. control schools, and also using an instrumental variables approach using as an instrument whether the student is zoned to attend one of the treatment schools.

While the effects overall for students in all grades were 0.276 SD in math and 0.059 SD in reading, they were remarkably larger in math specifically for 6th and 9th graders – equal to 0.48 and 0.74 SD, respectively. The fact that the largest impacts showed up exactly and only for the one subject and two grade levels that experienced Match tutoring leads to our conclusion that this is a promising intervention model. These gains equal two to four years of learning for the average American middle or high school student, according to data from the NAEP (Reardon, 2011). Because Fryer's findings come from a quasi-experimental study rather than a randomized experiment, they are inevitably subject to some uncertainty. But they are nevertheless striking.

Compared to regular classroom instruction, two-on-one tutoring greatly simplifies the instructional task that the adult is asked to carry out. Working with just two students makes it much easier for the instructor to individualize instruction (both in terms of the level and pace) to what students need. The tutoring method also makes it much easier to develop positive relationships with students, and to maximize time-on-task; one might

think of two-on-one tutoring as extreme class size reduction that would greatly reduce the risk of disruptions from other students (Lazear, 2001). Indeed because instructors basically do not need to worry about classroom management in this "teaching technology," the set of people who are capable of being effective tutors (in terms of either their abilities or prior training) is presumably much greater than the set of people who could succeed in teaching a large classroom of students. It is possible to provide such a high dosage (small student-to-tutor ratio, and high number of contact hours) by hiring recent college graduates, retirees, or career-switchers who are willing to do this for a year at modest pay as a public service. The intervention essentially substitutes a very different teaching method for many dimensions of what the previous literature has described as "teacher skill" or "quality" (such as teaching experience or extensive pedagogical training).

The two-pronged intervention that we study in this paper, outlined in the next section, addresses both the academic and non-academic barriers that we describe in this section. While our design does not allow us to separate out the effects of each prong of the intervention, the very large change in youth outcomes that we see as a result of the bundled intervention suggests that at least one of the two key barriers described here is indeed important in affecting how disadvantaged youth perform in school.

III. INTERVENTIONS

In this section we describe the bundle of academic and non-academic interventions that we study. This bundle of interventions was delivered in the context of an RCT that we carried out in a highly disadvantaged south-side Chicago high school during the 2012-13 academic year. Our research design, described more in section IV below, randomized some youth to receive just the non-academic intervention described here, other youth to receive both the non-academic intervention together with the academic intervention, and others to receive status quo services. In practice there appear to have been some spillover effects from the academic intervention to those youth who were assigned to only receive the non-academic intervention, as we discuss in more detail in section VII. This complicates our ability to isolate the effects of the two different prongs of the intervention we study here.

A. Non-academic intervention

The non-academic portion of the intervention we study here, called "Becoming a Man" (BAM), was developed and implemented by a Chicago-area non-profit organization, Youth Guidance (YG). It includes in-school programming that exposes youth to pro-social adults, and provides them with social-cognitive skill training that follows the principles of cognitive behavioral therapy (CBT).

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

The BAM program includes a mix of elements. Some of the BAM curriculum consists of what might be called character or values education. The curriculum also includes efforts to develop specific social or social-cognitive skills such as generating new solutions to problems, learning new ways of behaving, considering another's perspective, thinking ahead, and evaluating consequences ahead of time. This is similar to what many social-emotional learning (SEL) programs try to achieve.

The BAM curriculum also includes standard elements of CBT (Beck, 2011), which tries to address the problem that a great deal of behavior stems from automatic decision-making (what psychologists call "System 1" behavior), but people are rarely aware of either the automatic thoughts that drive their behavior or the predictable biases to which automatic thoughts are prone. For example, people often overgeneralize and assume that a single negative event is symptomatic of a broader problem, or "catastrophize" and make a negative event more negative than it is. Or people jump to conclusions, forming negative interpretations even before there is evidence to support them. CBT addresses these problems by making people more aware of their own thoughts and how their thoughts drive behavior. CBT as a psychologist would deliver it would often be one-on-one; the BAM intervention is delivered in small to medium sized groups (10-15 youth per group) to hold down costs.

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

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

B. Academic intervention

The academic intervention included in our RCT was delivered by staff hired by our own research team but was modeled closely on the Match model. We selected the Match tutoring model as the academic component of our two-pronged intervention because of our hypothesis that "academic mismatch" is an

important problem in many urban high schools and that intensive individualized instruction is a promising solution. Bloom (1984) summarizes a series of RCTs with elementary and middle school students teaching them new subjects about which they would have little prior background (cartography and probability), and found that students assigned to receive one-on-one or small-group (not more than three-on-one) tutoring generated average test scores that were fully two standard deviations higher than those of students assigned to regular classroom instruction. Compared to regular classroom instruction, tutoring also generates large increases in time-on-task (90+% versus 65%) and improved student attitudes and interests. Tutoring by its nature was found to increase the amount of feedback and correction between student and instructor, a key characteristic of effective teaching, and also ensures that all students receive this attention – including those students who are struggling in school. There is some indication in these studies that teachers in regular classrooms tend to focus their attention on students in the top third of the achievement distribution. The challenge for education policy has been that such intensive small-group tutoring is very costly. The "two sigma problem," as Bloom described it, is to identify lower-cost instructional alternatives that can be as effective as tutoring.

One major innovation of the Match model, and a key reason we selected it, is the recognition that the instructional "technology" of tutoring is quite different from that of a classroom and so the set of skills and experiences required to be a successful instructor are different. This enables Match to expand the pool of people to recruit to be tutors and focus on people who are talented with strong math skills and willing to devote a year to public service, but who do not necessarily have extensive prior training or experience as teachers. As with other public service programs like Teach for America or City Year, the tutors were willing to work at relatively low wages ($16,000 plus benefits for the nine-month academic year). This makes the incredibly high dosage of the Match tutoring model feasible.

Another reason we selected the Match model was because of the promising findings from Fryer's (2011) non-experimental study. We used our own staff to try to implement the Match model as faithfully as possible during the pilot results reported on here, rather than subcontract with Match, because of other obligations that Match Education had for our pilot year (AY2012-13). (Our research team is now currently in the field with a large-scale experiment in which Match Education itself is working under sub-contract with our research team to

provide their tutoring intervention. Compared to the pilot results that we report on here, our ongoing large-scale experiment will also provide for a cleaner test to separate out the effects of BAM versus Match tutoring).

During the school day, students as part of their regular class schedule were assigned to participate in a one-hour-long tutoring session, every day. Each tutor worked with two students at a time during each session. The tutors were mostly recent college graduates who were hired because they have very strong math skills and interpersonal skills, although as noted above they did not have formal teacher training and were not licensed Illinois teachers. The program in its essence thus shares many similarities with that of Banerjee et al. (2007).

The Match intervention individualizes instruction but, unlike tracking, is adaptive: rather than locking students into a particular instructional level, Match tailors the level of material to students' changing needs and allows them to progress as quickly as they are able to learn. Tutoring sessions have a curriculum, with about half of each session devoted to working on subject material that students are working on in class, targeted to the Illinois state Common Core standards. (The school supervisor for the tutoring intervention coordinated with math teachers in the school to obtain ahead of time their weekly lesson plans.) We used a commercial curriculum to provide remedial skill development for the other half of each session, together with frequent formative assessments and curriculum adjustments in response to the assessments.

The control group in our study was eligible for all of the other academic supports currently available in the high school. Every 9th grader in our study sample (those assigned to treatment as well as to our control condition) was receiving a double period of math. Control students remained eligible for the tutoring that is provided by CPS with No Child Left Behind (NCLB) funding, which provides 21 hours of writing tutoring and 20 hours of math tutoring *per year* (or about ½ hour *per week* of math tutoring, compared to about one hour *per day* with our Match-style academic intervention for the treatment group).

IV. DATA

One source of data we have for measuring "dosage" is provider records. Youth Guidance shared with us individual-level records on participation in the weekly BAM sessions. We also have daily logs from the tutoring team that records attendance by youth assigned to receive the academic intervention as well.

Our main source of both baseline information about youth and their subsequent outcomes comes from longitudinal student-level records maintained by CPS. Because our study sample was initially drawn from students attending our study school, we have CPS student ID numbers for everyone we randomly assigned. So our initial match rate to the CPS administrative records for our study sample is 100% by construction.

From CPS we obtained student-level school records for the academic years 2011-12 (the year before our intervention was fielded) and 2012-13 (the intervention year itself). These CPS student records include whether the student has a disability (as indicated by having an individualized educational plan, or IEP; all but one of the students in our study sample who had an IEP were classified as "learning disabled"); month and year of birth (so we can construct age); race / ethnicity; eligibility for free and reduced price lunch; course grades in each subject (so that we can examine impacts on grades in specific subjects such as math or on overall GPA); and enrollment status, so that we can examine dropout versus school persistence.

These data also include achievement test scores for the exams that CPS administers to 9th and 10th graders – the 9th grade EXPLORE and 10th grade PLAN tests, which are developed by ACT, Inc. The EXPLORE exams include a 40-item, 30-minute English test; a 30-item, 30-minute reading test; and, particularly relevant for our purposes, a 30-item, 30-minute math test, which, as ACT notes, covers "four areas – knowledge and skills, direct application, understanding concepts, and integrating your understanding of concepts," in pre-algebra (10 test items), elementary algebra (9 items), geometry (7 items), and statistics and probability (4 items).[9] The 10th grade PLAN tests include a 30-minute English exam (30 items on usage/mechanics, 20 items on rhetorical skills); a 20-minute, 25-item reading exam; and a 40-minute math test that covers pre-algebra and first-year algebra (22 items), and plane geometry (18 items).[10]

The EXPLORE and PLAN tests provide results both as scaled scores, and in terms of the student's percentile rank within the national distribution of test takers. We pool together results for 9th and 10th graders and report impact estimates using the test score results scored in three different ways. First, we show test score

[9] See http://www.act.org/explorestudent/tests/math.html. Sample problems from the EXPLORE 9th grade math test are available at: http://www.act.org/explorestudent/pdf/math.pdf

[10] See http://www.act.org/planstudent/tests/index.html. Sample problems from the PLAN 10th grade math test are available at: http://www.act.org/planstudent/pdf/sample.pdf

results using the EXPLORE and PLAN scale scores normalized to our control group's distribution, that is, subtracting off the control mean from each student's score and then dividing by the control group's standard deviation. This is the convention that is widely used in education research, known as Glass's Δ (Glass, 1976), so reporting our test score results scaled in this way has the advantage of facilitating comparisons to other studies. Second, we present test score results that standardize the scale scores using the national distribution for the scale scores. Third, we present test score results in terms of national percentile rankings. The last two metrics have the advantage of letting readers see how the intervention moves children within the national distribution.

We can also calculate the CPS "on track" indicator that was developed by the Chicago Consortium on School Research (CCSR) for high school freshman. Students are on-track if they accumulate five full-year course credits (in any credit-bearing class), and accumulate not more than one semester F grade in a core class (Allensworth and Easton, 2005). Since our study sample consists of both 9th and 10th graders, we extend the basic logic of the "on track" indicator and calculate the measure for students in both grades that we study.

For this study we were unable to match youth to government arrest records, so we cannot compare the size of any effects on criminal behavior from this intervention cocktail with the effects of providing youth with BAM alone, as reported in Heller et al. (2013).

V. STUDY SAMPLE AND RANDOM ASSIGNMENT

The main challenge with any intervention study is the possibility that the youth who wind up receiving programming are systematically different from those who do not. Our study overcomes that challenge through random assignment of eligible youth to one of two programming conditions (either BAM or BAM plus Match-style high-dosage tutoring), or to a control group. By virtue of random assignment, we would expect the youth assigned to our three groups to be the same in expectation.

Our research team carried out the random assignment for the current project ourselves. In October 2012 (the fall of our intervention year), we used CPS administrative data to identify male youth who were in either 9th or 10th grade and enrolled at our study high school. Because we are studying a school-based intervention, we excluded youth who in the previous academic year (AY2011-12) missed more than 60% of all school days and

failed more than 75% of their classes, with the logic that they would be unlikely to attend school enough during our intervention year (AY2012-13) to benefit.

For the remaining male students, we calculated an "academic risk index" that is a function of the number of prior-year course failures, unexcused absences, and being old for grade (previously held back). We then ranked students on the basis of this risk index, and selected 106 male 9th and 10th graders with the highest risk scores to be in our study sample. Given that the study high school's total enrollment is slightly less than 1,000, our study sample of N=106 male youth represents about one-third of all males in 9th and 10th grade in the school. These youth were randomly assigned to one of three conditions:

(1) *Control (N=34)*;

(2) *BAM only (N=24)*;

(3) *BAM plus Match-style high-dosage tutoring (N=48).*

Because we had different levels of capacity for the tutoring and BAM within the school, our random assignment algorithm intentionally over-assigned eligible youth to the group that received both our academic and non-academic intervention, with a lower assignment probability for the non-academic-only group.

For those assigned to programming, consents for program participation were sought from youth and their parents. Our team was able to access administrative data on youth assigned to all three groups, including youth assigned to the control group, and youth assigned to the treatment groups who chose not to participate. All of our study procedures were approved by the University of Chicago IRB.

Table 1 shows that the average baseline characteristics are generally quite similar across randomized groups, which is what we would expect with properly executed random assignment. Because of limits to our statistical power, most of the results we report below compare all youth assigned to treatment (pooling youth assigned to BAM with those assigned to receive BAM plus Match-style tutoring) with all youth assigned to control. None of the pair-wise comparisons of baseline characteristics are statistically significant.

If we instead compare the average baseline characteristics for youth assigned to the control group (N=34) to those youth assigned specifically to receive BAM plus tutoring (N=48), none of the pair-wise differences are statistically significant. If we compare the control group to those assigned to the BAM-only

group (N=24), we see one pair-wise difference that is statistically significant at the $p<0.05$ threshold (disciplinary incidents during the previous year, i.e. AY2011-12, of 0.79 incidents over the year – which is large relative to the control mean of 1.82) and one difference that is statistically significant at the $p<0.10$ level (fall 2012 reading scores on the EXPLORE / PLAN tests that were administered prior to randomization; the difference is about four percentile points in the national distribution). A formal baseline balance test that considers the full set of baseline characteristics simultaneously cannot reject the null hypothesis that the distributions of all baseline variables together are jointly the same (Appendix Table 1).

Table 1 shows that our study sample is entirely male, and almost entirely African-American. The rest of the table highlights the high level of disadvantage for this study sample and their significant academic challenges. All but one of the youth in our study sample (99%) are eligible for free or reduced price lunch (fully 94% are eligible for free-lunch specifically). Over one-quarter of our study sample has an individualized education plan (IEP), with most of these diagnosing some sort of learning disability. During the previous year (AY2011-12), male youth in our study sample missed an average of 19.5 days of school, experienced 3.2 out-of-school suspensions, failed 1.7 classes, and had an overall GPA of 2.15 on a four-point scale. During the fall of 2012, before we began providing intervention services, the average youth scored at the 26^{th} percentile of the national distribution on their EXPLORE / PLAN reading test and at the 22^{nd} percentile in math.

VI. ANALYSIS PLAN

Given the experimental design of our study, our analysis plan is quite straightforward. Our estimating equation for the effect of being offered programming – the intention to treat effect (ITT) – is given by equation (1). Let Z_i represent treatment assignment, either a vector of two indicators for assignment to the BAM and BAM-tutoring treatment arms, or (to improve statistical power given our small sample) a single indicator capturing assignment to either treatment group. Let Y_{it} represent some outcome of interest during the post-randomization period (t), let B_i be a "randomization block" indicator (effectively a grade-10 indicator), and let $X_{i(t-1)}$ be a set of pre-randomization baseline characteristics that include prior reading and math achievement test scores, IEP status, previous year GPA, absences, suspensions and disciplinary incidents, and socio-demographic characteristics (age, grade and free lunch eligibility). We include these baseline characteristics to help account

for residual variation in the outcome of interest, thereby improving the precision of our impact estimates, although the results are qualitatively similar without these controls.

(1) $Y_{it} = \pi_0 + Z_i\pi_1 + X_{is(t-1)}\pi_3 + \pi_4 B_i + \varepsilon_{it}$

The random assignment of youth to treatment or control conditions ensures that under standard assumptions estimation of this model by ordinary least squares will yield unbiased and valid estimates of π_1. To ensure that the standard errors we calculate are not misleadingly small as an artifact of the modest number of youth in our study sample, we also report p-values that come from a non-parametric permutation test (Efron and Tibshirani, 1993). These are calculated by randomly re-assigning values of the treatment indicator across our sample 100,000 times, and calculating the t-test statistic for the placebo treatment versus control contrast in each replication. The permutation test p-value is the share of replications where the t-test statistic exceeds the value that we calculate using the actual treatment assignment variable.[11]

The main threat to valid inference comes from selective sample attrition, and indeed Table 2 shows that youth assigned to any treatment (third column) turn out to be about eight percentage points more likely than controls to have valid scores for the spring 2013 (post-random assignment) EXPLORE and PLAN achievement tests. The results that we present below suggest the differential rate of missing-ness for end-of-year test scores does not seem to be due to treatment effects that reduce school dropout, but could be due to treatment effects that reduce student school absences.

In either case, we would expect relatively weaker students to be more likely to be missing tests, which Table 3 suggests is indeed the case. Comparing columns 1 and 2 of Table 3 we see that students assigned to the control group who are missing end-of-year 2013 (post-randomization) test scores have lower prior-year grades and achievement test scores than those who have valid test results, and also have higher rates of absences, suspensions, and IEP designations. Columns 3 and 4 of Table 3 show the same is true for youth assigned to the treatment group for whom we do versus do not have valid spring 2013 achievement test scores. If the treatment serves to increase the rate at which more academically marginal students take the end-of-year spring 2013 (post-

[11] For the permutation tests for the effects of treatment on the treated (TOT), described below, we randomly re-assign both the endogenous variable for actual treatment participation (D) and treatment assignment (Z).

randomization) test, then a comparison of the average of the valid test scores for the treatment versus control groups should understate the true beneficial effect of the intervention on academic achievement.

We also empirically explore the sensitivity of our results to different methods for dealing with selective attrition. Showing the sensitivity of the results to inclusion or exclusion of baseline covariates is one test of whether youth with missing scores have different values of baseline characteristics. In addition we present results that impute missing spring test score results using multiple imputation (MI), which has the limitation of assuming that these outcome data are conditionally missing at random (MAR) – that is, conditional on observable baseline characteristics, missing-ness is unrelated to unobserved attributes of students. Finally, we present results from quantile regressions that focus on estimating treatment-control differences in the median test score rather than the mean, and impute arbitrarily low test-score values to those with missing scores. This approach requires the assumption that students missing post-test data have actual post-test scores that fall in the bottom half of the sample distribution.

While the ITT takes full advantage of the experimental design, because not all youth offered programming choose to participate, the ITT will understate the effects of actually receiving services. We therefore also report the effect of actually participating in treatment – the treatment on the treated (TOT) effect – by using random assignment (Z_i) as an instrumental variable (IV) for participation (D_i), as in equations (2) and (3) (Angrist, Imbens & Rubin 1996; Bloom 1984). The IV estimate for the parameter β_1 in equation (3) is essentially a ratio of two ITT estimates – the ITT effect on the outcome of interest in the numerator, with the ITT effect on program participation rates in the denominator. With a participation rate of 74% for all youth assigned into either treatment arm, the TOT estimate will be about 1.35 times the ITT.[12]

(2) $$D_i = \gamma_0 + Z_i\gamma_1 + X_{is(t-1)}\gamma_2 + \gamma_3 B_i + \mu_i$$

[12] The participation rate for the 24 youth assigned to the BAM-only treatment arm was 71%, and for the 48 youth assigned to the BAM+Match tutoring treatment arm equaled 75%. We define participation as "attended at least one program session." Youth assigned to the BAM+Match treatment are counted as a participant if they attend at least 1 session of either program. Some readers might think this is a low bar for defining what counts as "participation." Note that our approach is conservative in the sense that using a higher threshold for participation (that is, counting only youth who attend some higher number of sessions as participants) would have the result of further increasing the size of our TOT estimates, by essentially allocating the ITT effect over fewer youth. But using a higher participation threshold boosts the TOT estimate by assuming that youth who participate in fewer than N sessions do not benefit at all from participating in those sessions, which seems like a strong assumption since so little is currently known about the functional form of the treatment dosage / treatment response relationship for these types of programs.

(3) $Y_{it} = \beta_0 + D_i\beta_1 + X_{is(t-1)}\beta_2 + \beta_3 B_i + \upsilon_{it}$

Note that we are *not* estimating the effects of program participation by comparing participants to non-participants; that sort of *non-experimental* estimate would likely be biased by the fact that program participants and non-participants are different on average (see Appendix Table 2). The IV estimate is nearly fully experimental; we say "nearly" because the IV estimate requires for unbiased estimation the same assumption as does the ITT estimate (that randomization was carried out correctly), but now adds one more assumption – that treatment-group assignment has no effect on the behavior of youth who do not participate in the intervention.

Because none of the youth assigned to our control group received services, our IV estimate for β_1 represents an estimate for the TOT rather than a local average treatment effect (LATE). If youth vary in how they respond to or benefit from program participation, then our TOT estimate does not capture the average effect that would result if everyone participated. Nevertheless the TOT estimate is still an interesting parameter, because it tells us something about the average effect we might expect if we were to deliver this intervention to similar sorts of schools to the one we study here, and if a similar type of youth were to participate. Another advantage of the TOT is that it facilitates comparison of our effect sizes to those of other studies.

To benchmark the size of the TOT effect, we present the control complier mean (CCM), that is the average outcome for those in the control group who would have participated in programming had they been offered the chance, calculated as in Katz, Kling and Liebman (2001). The CCM can differ from the overall control mean if the type of person who would participate in programming if assigned to treatment is systematically different from the non-participants among those assigned to treatment.

We examine three main outcome domains or "families" of outcomes, with three outcomes per family:

(1) *Math achievement* (performance on the 9th and 10th grade EXPLORE and PLAN mathematics tests, scored in different ways; math grade point average, or GPA; and math course failures);

(2) *Other (non-math) academic achievement* (performance on the EXPLORE and PLAN reading test; non-math GPA; and non-math course failures;

(3) *Behavior* (absences, school student misconducts, and out of school suspensions).

We also present results on the CPS "on track" indicator for high school graduation, which is essentially a summary index that combines different elements of our math and non-math achievement domains.

In our main results tables we present the results of three types of statistical tests: standard t-test statistics and p-values that come from considering just the single pairwise comparison of treatment and control for the particular outcome being examined in that regression; p-values that account for multiple testing concerns by controlling for the family-wise error rate (FWER), or the probability that at least one of the true null hypotheses in a family of hypothesis tests is rejected, using a free step-down resampling method[13]; and the false discovery rate (FDR) within each family, or the proportion of null-hypothesis rejections within a family that are type I errors or "false positives" (Anderson, 2008), calculated using the two-step procedure from Benjamini, Krieger and Yekutieli (2006). (FDR q-values calculated using the one-step procedure from Benjamini and Hochberg 1995 are similar.) Because the TOT is basically just a re-scaled version of the ITT, with a similar scaling factor for the point estimate and standard error, the t-statistics are very similar (and in a model specification without covariates would be the same). To simplify our tables we report the p-values calculated in different ways for the ITT only; the p-values for the TOT are always very similar, and are reported separately in the appendix.

We believe there is a case to be made to focus on the FDR values in our tables; ours is a version of the standard "multiple end points" problem when comparing some alternative program to status quo (see Benjamini and Hochberg, 1995), where the decision about whether to prefer the alternative to status quo will depend on the set of outcomes being compared across the two states rather than any individual outcome. FWER control is more conservative for this purpose than FDR control, while pairwise comparisons are not conservative enough. In any case we report the results from each method for completeness.

---

[13] Specifically, we use a bootstrap resampling technique that simulates data under the null hypothesis (Westfall & Young 1993). Within each permutation, we randomly re-assign treatment and control indicators with replacement and estimate program impacts on all three outcomes within each of our outcome domains (we do this separately for each domain). By repeating this procedure 100,000 times, we create an empirical distribution of t-statistics that allows us to compare the actual set of t-statistics we find to what we would have found by chance under the null. We maintain the original sampling frame for each iteration, assigning the same number of pseudo-treatment and pseudo-control youth as in our original sample. This technique preserves the correlational structure and underlying distributions of our data, providing the adjusted probability we would observe our results by chance given our data and the number of tests we run. Rather than use a single p-value adjustment for all the outcome measures, we use a free step-down procedure to adjust the p-value on each outcome separately. The idea is that once a null hypothesis has been rejected via the bootstrap resampling method, it is removed from the family of hypotheses being tested (thus increasing the power of the remaining tests). We then calculate a new adjusted p-value with the bootstrapped empirical distribution of t-statistics for only the remaining tests, providing a more powerful adjustment than setting all p-values to the same minimum value.

VII. MAIN RESULTS

In this section we begin by documenting the "dosage" of the programming that youth received in our study, followed by a discussion of our estimates for the impacts on student school outcomes. The size of the impacts from this intervention on schooling outcomes is quite large, particularly given how relatively modest the treatment dosage is for so many of the youth in our study sample.

A. Program participation rates and "crossover"

As noted above, the participation rate among all youth assigned to receive programming (pooling the youth assigned to BAM only and those assigned to receive BAM plus match tutoring) was 74%. The participation rate for the 24 youth assigned to the BAM-only treatment arm was 71%, and for the 48 youth assigned to the BAM+Match tutoring treatment arm equaled 75%. (Of those assigned to BAM+Match, 36/48, or 75%, ever attended at least one BAM session while 19/48 or 40% participated in tutoring).

In our main analyses we initially define participation as "attended at least one program session." Youth assigned to the BAM+Match treatment are counted as a participant if they attend at least one session of either program.[14] Figure 1 presents data for those youth who participated in at least one BAM session, showing the frequency of sessions attended for those assigned to the BAM-only group and separately for those assigned to the BAM+Match group. Figure 2 presents the distribution for number of tutoring sessions attended among those who were assigned to the BAM+Match group and ever attended at least one tutoring session.

It turns out that there was some informal treatment crossover between the BAM-only and BAM+Match tutoring groups that is not captured by our administrative records on tutoring participation, which complicates the interpretation of the effects of the BAM-only treatment that we present below. The BAM sessions were held in the same empty classroom within our study high school as housed the tutoring sessions. While the BAM group sessions and the tutoring were always held at different times, the tutors report that youth assigned to receive BAM programming who were being too disruptive in class were often sent by teachers to what became

[14] Some readers might think this is a low bar for defining what counts as "participation." Note that our approach is conservative in the sense that using a higher threshold for participation (that is, counting only youth who attend some higher number of sessions as participants) would have the result of further increasing the size of our TOT estimates, by essentially allocating the ITT effect over fewer youth. But using a higher participation threshold boosts the TOT estimate by assuming that youth who participate in fewer than N sessions do not benefit at all from participating in those sessions, which seems like a strong assumption since so little is currently known about the functional form of the treatment dosage / treatment response relationship for these types of programs.

known as the "BAM room" to "cool down." If the tutors were there in the room at the same time and a tutor had excess capacity, he or she would work with that youth. (Unlike with the structured Match-style tutoring, which always focused on math, this sort of informal tutoring occurred on a range of subject areas.)

There might have also been spillover across treatment arms through two other mechanisms as well. First, the structured BAM groups were composed of some youth assigned to BAM only and some youth assigned to BAM+Match tutoring. This could have led to some peer spillover effects on math achievement from the BAM+Match treatment to those assigned to BAM only. The tutors also report that there may have been a spillover effect that operates through changes in teacher expectations. Many of the youth in our study sample were very far behind. Youth assigned to BAM+Match tutoring did much better in school, particularly in math class, as we show below. The tutors believe that seeing these gains among BAM+Match youth changed math teacher expectations for these students and the patience that teachers had when interacting with these youth, which could have spilled over to other youth assigned to the BAM-only group.

In our results section below we provide some imperfect but suggestive empirical tests that explore the plausibility and empirical importance of these candidate sources of spillover across treatment arms.

B. Impacts on student learning and behavior

Table 4 presents our main findings for the effects of being offered the chance to participate in our two-pronged intervention (intention to treat, or ITT) and the effects of actually participating (the effects of treatment on the treated, or TOT). As a benchmark to help interpret the size of these effects, we also report the control mean (CM) and the control complier mean (CCM). We initially pool both treatment arms together to improve statistical power – that is, we initially show the results from comparing all youth who get either BAM or BAM+Match versus the control group. However, we also report in a later table the results of considering the BAM-only and BAM+Match groups separately.

The first "family" of results shown in Table 4 is math achievement, where we see very large gains from the intervention. The first row shows the ITT effect on spring math test scores reported as Z-scores that are standardized using the control group distribution, which is reported in the same metric as in most other education studies (Glass' Δ, i.e. subtracting off the control mean and dividing by the control group's standard

deviation). The ITT measured this way is 0.51 SD; focusing just on this outcome variable (not adjusting for multiple comparisons) the per comparison error rate (PCER) using the regular OLS standard errors equals p=0.013, and equals p=0.016 when we instead use a permutation test with 100,000 replications. When we account for multiple comparisons and control the family-wise error rate (FWER), that is, the probability of at least one false positive result within this family of three math-achievement outcomes, our p-value equals 0.036.[15] The false discovery rate (FDR) q-rate equals 0.033. That is, the result is significant if we are willing to accept that about one of every 30 statistically significant point estimates is actually a false positive. The TOT effect on this fairly broad national test of math achievement is 0.65 SD.

The advantage of reporting our math test-score result in terms of Glass' Δ is we can compare our result to other studies that use the same test-score scaling, although expressing the effect in terms of how far students move in the national distribution is also of interest. The TOT effect in the national distribution (using the math scale scores) equals 0.48 SD, or about 60 percent of the black-white test score gap in the national NAEP test for 13 year olds (see Figure 3). As we would expect, the p-values are nearly identical to those we see when we standardize these math scores using the control group distribution rather than the national distribution.

Perhaps more intuitive is to think about these impacts in terms of exactly where the student falls in the national distribution. The sort of youth in the control group who would have participated in programming if offered the chance (the CCM) has a spring math score that falls at about the 19th percentile of the national distribution on the EXPLORE and PLAN tests. Participating in programming (the TOT effect) increases the ranking in the national distribution by nearly 15 percentile points, to about the 34th percentile.

The top panel also shows a large gain in math GPA over the course of the academic year, with a TOT effect equal to 0.58 points on a 4-point GPA scale, compared to a control complier mean of 1.24 math GPA. Put differently, the average control complier youth gets about a D in their math class while the average treatment-group complier scored at about a C (this result is statistically significant at the usual 5% threshold regardless of whether we focus on the PCER, the FWER, or the FDR). This effect equals 0.67 SD in the control group's math

[15] We say "three variables" even though there are more than three variables technically reported in this panel because the different math test score and math GPA variables are just re-scaled versions of the same underlying thing.

GPA distribution. The next row shows that the likelihood of failing math during our intervention year is reduced by about two-thirds (TOT effect of 0.42) which is quite large as a share of the CCM (0.68) but given the sizable standard error just statistically significant at the p<.10 level even with just the pairwise test.

The second panel of Table 4 shows there are no statistically significant spillover effects on reading test scores or GPA outside of math classes, although we do see a sizable reduction in course failures in classes outside of math. The control complier mean is 3.54 course failures over the course of the 2012-13 academic year, while the effect of participating in programming (the TOT) equals a reduction of two course failures. This effect is statistically significant at p<.05 regardless of how we calculate our p-value.[16]

The third panel of Table 4 shows that there are sizable (albeit not quite statistically significant) impacts on two of our three measures of behavior. The CCM for absences is equal to 45, or an average of nine weeks of school missed over the 2012-13 academic year. The effect of participating (TOT) is equal to 12.9 fewer days missed, or just over one-quarter. This is statistically significant with just the pairwise t-test but not quite significant when we adjust for multiple comparisons; the FWER controlled p-value = 0.109, while the FDR q-value = 0.141. The effect on out of school suspensions is equal to one half of the CCM but, given our sizable standard error, is not statistically significant.

Figure 4 shows that program participation increases the likelihood that a student is "on track" (according to a CPS indicator that summarizes a number of the math and non-math achievement variables) by nearly one-half. The control complier mean (CCM) for 9th and 10th graders to be "on track" for graduation is equal to 53%. Program participation (the TOT effect) increases that likelihood by fully 24 percentage points, or 46% of the CCM. Compared to students who are not "on track" according to this indicator, those who are on track have four-year graduation rates that are 59 percentage points higher, and five-year graduation rates that are 57 percentage points higher (Allensworth and Easton, 2005, p. 8). If we take this correlation between the on-track indicator and high school graduation rates at face value, our estimated effect on the on-track indicator increases

[16] Table 4 reports a FDR q-value for failures in non-math subjects that is slightly higher (0.022) than the FWER (0.019), while conceptually we expect the FDR to never be higher than the FWER. This can happen sometimes for the most significant outcome within a family or outcome domain because of the way the FDR is calculated in practice. This is easiest to see within the context of the one-step FDR procedure from Benjamini and Hochberg (1995), which shows that for the most significant outcome within a family or outcome domain the FDR calculation can sometimes collapse to the Bonferroni calculation, which is conservative.

expected high-school graduation rates by around 14 percentage points (nearly half of the control group's expected 30 percent graduation rate). Thinking about the statistical significance of the on-track indicator within our multiple-testing framework is a bit complicated because it is an index of outcomes that span outcome domains (math and non-math achievement), but the estimated effect on the on-track indicator is at least significant using the pairwise comparison ($p<0.05$).

C. Sensitivity analyses

Table 5 shows that in general our results are qualitatively similar regardless of whether or how we control for baseline covariates. Since the p-values are nearly identical for the ITT and TOT estimates for a given outcome and model specification, we present just the ITT effects to simplify the table. The first panel shows that when we use math achievement test scores as our outcome of interest, not controlling for baseline covariates at all, as shown in the last column, tends to reduce the t-statistics somewhat by making the point estimates about one-quarter smaller than in our preferred model specification and makes the standard errors about one-quarter larger. The magnitude of the ITT effect is still quite large even without covariates (0.41 SD for the version of our test score outcome that is normalized using the control group distribution, and 0.30 SD in the national distribution). The remaining panels show the results for other outcomes for which we have fewer missing observations tend to be less sensitive to whether or how we control for baseline covariates, which provides an indication that spring 2013 (post-randomization) math scores are not missing completely at random.

Table 6 shows what happens when we account for missing follow-up data in different ways, focusing again on the ITT for simplicity. Missing data is mostly an issue for our standardized test-score measures, where we are missing valid spring 2013 (post-randomization) scores for about one quarter of our overall study sample (25 out of 106 youth). Compared to our main results, the point estimates for our math test results are slightly smaller when we use multiple imputation to fill in missing values to the spring achievement test scores, and the standard errors are larger, but the point estimate itself still suggests a sizable change in math scores. Assigning arbitrarily low scores to missing values and using quantile regression implies an effect on the median score of 0.39 SD (using the control group's distribution), compared to an effect on the median of 0.42 when we use just non-missing observations.

The two other relevant outcomes besides tests scores for which missing data are an issue are for days absent from school and for number of out-of-school suspensions (shown in the bottom panel of Table 6). For both of these variables we are missing a total of eight observations out of our sample of 106 youth. We see that the results for the treatment-control difference in the mean are quite similar when we use multiple imputation to fill in missing values, or assign arbitrarily low values and use quantile regression to estimate treatment effects on the median values of these outcomes.

D. Results by separate treatment group

Given the relatively modest number of youth in our study sample, our main analyses have so far pooled together youth assigned to the BAM-only group and those assigned to BAM+Match tutoring into a single "treatment group" to be compared to control-group youth. In this section we report the results for the two treatment arms separately, although interpretation of these results is complicated somewhat by the fact that there may be some treatment crossover or spillovers between the two groups as described above.

Figure 5 presents the TOT point estimates and 95% confidence intervals for the separate effects of the BAM-only treatment and BAM+Match tutoring. We focus on comparing the TOT effects of the two interventions, to avoid confounding differences in the effects of the programming actually received with differences in the program take-up rates. The estimated TOT point estimates are usually slightly larger for the BAM treatment arm than for the BAM+Match treatment arm. Given our sample sizes, the confidence intervals for both TOT effects are quite large, with a great deal of overlap with one another. Using a permutation test for statistical inference, in no case can we reject at the usual 5% threshold the null hypothesis that the TOT effects are the same (Table 7).

While our confidence intervals are too wide to say anything about which of the two intervention arms is more effective, the fact that the point estimate for the BAM-only effect on academic outcomes like math test scores is so large seems to stand in contrast to the results of our team's last study of BAM, where we saw no detectable effects on achievement test scores (Heller et al., 2013).

Tables 8 and 9 explore the hypothesis that there may have been some spillover between the tutoring intervention and the BAM-only group when youth in the latter were kicked out of class and sent to the room

where BAM sessions and tutoring occurred to "cool off." The tutors report that in these circumstances they would, if they had some slack time, work with those youth on their schoolwork. This hypothesis suggests that we should see the biggest effect of assignment to the BAM-only group on academic outcomes for those youth who are most likely to be disruptive and get kicked out of their regular classrooms.

Table 8 provides some tentative support for this hypothesis. Each cell of the table reports the results of running a regression for the outcome defined in the leftmost column (so each row is an outcome), using just the sample of youth assigned to either control or the BAM-only treatment, and interacting treatment assignment with the number of disciplinary actions or out-of-school suspensions the youth has either during the pre-program year (AY2011-12) or during the program year (AY2012-13). The first panel shows that youth who are more likely to get into trouble in school have *higher* math test scores than those youth who do not get into trouble (that is, the interaction between our measures of disciplinary problems and treatment assignment are positive). Given our small sample sizes, these interaction terms are not statistically significant, but they imply substantively large differences in math scores. For example, two youth assigned to the BAM-only group whose number of disciplinary incidents during the pre-program year (AY2011-12) differed by one standard deviation (about 2.1) would experience gains from BAM-only assignment of 0.49 versus 0.86 SD, respectively.

We see a qualitatively similar pattern if we use disciplinary actions or out-of-school suspensions during the program year itself (AY2012-13), as seen in the last two columns. Recognizing that our standard errors are sizable and that this analysis is non-experimental (that is, we are interacting a post-randomization measure with treatment assignment), the data suggest that youth who are getting into trouble more even during the program year itself may be experiencing larger gains in math scores than other youth.

The rest of Table 8 shows that those youth who get into trouble relatively more frequently in school seem to if anything benefit relatively less from treatment assignment with respect to outcomes like grades; this is what we would expect if those measures are a function of student behavior or demeanor as well as learning. Table 9 shows that the pattern is qualitatively different for youth assigned to the BAM+Match group. Perhaps most revealingly, while there is some indication that youth who were getting into the most trouble during the pre-program year (AY2011-12) benefited more from BAM+Match with respect to many of our outcome

measures (e.g. math grades, non-math grades), this is not true for math test scores, and it is also not true when we use measures of getting into trouble during the program year itself (AY2012-13).[17]

VIII. CONCLUSIONS

The conventional wisdom around efforts to help disadvantaged youth is nicely summarized by Barrow, Claessens and Schanzenbach (2013): "The finding of no test score improvement but a strong improvement in school attainment is consistent with a growing literature suggesting that interventions aimed at older children are more effective at improving their non-cognitive skills than their cognitive skills." This less-than-stellar track record of previous efforts has led to calls for re-orienting high schools for disadvantaged youth to focus more on vocational or technical training (Cullen et al., 2013), or for policymakers to focus more resources on academic interventions in early childhood instead (for example Carniero and Heckman, 2003).

The impacts of the pilot intervention reported on in this paper are large enough to raise the question of whether the field has given up prematurely on the possibility of improving academic outcomes for disadvantaged youth. Our hypothesis is that a systemic problem in many current urban schools is the lack of a sufficiently intensive safety net to remediate academic or non-academic barriers to youth engaging with classroom-level instruction, which leads to "mismatch" between what many students who are falling behind need and what regular school settings deliver. Have previous interventions mostly aimed at the wrong target?

We find that a two-pronged intervention that addresses the maladaptive automatic behaviors that impede youth from successfully engaging with school and, at least as importantly, better individualizes the academic instruction that disadvantaged youth receive, seems to generate large gains in learning with a sample of low-income male high school students living in a very distressed urban area. Participation reduces course failures by about 66% in both math and non-math classes, increases rates of being "on track" for graduation (and hence expected high school graduation rates) by nearly one-half, and shows large gains in a broad measure of math

[17] Another test we carried out is to examine whether youth in the BAM-only group who are assigned to a BAM group with a relatively higher share of tutoring participants from the BAM+Match tutoring group experience relatively larger gains in math scores. This does not seem to be the case, but the number of youth contributing data to this analysis is very small. The comparison is also complicated by the fact that which BAM group a youth is assigned to is a function of their schedule for AY2012-13, which is endogenous.

test scores equal to 0.48 standard deviations in the nationwide test-score distribution, and 0.65 standard deviations using the control group distribution (the way most education studies report results).

As one way to judge the magnitude of our test score result, the effect measured relative to the nationwide test-score distribution is equal to about 60% of the black-white test score gap in math in the National Assessment of Educational Progress (NAEP) among 13 year olds (which equals 0.80 SD). This does not mean that providing this intervention universally would cut the black-white test score gap by 60%, since the effects could be different for different populations and in particular we have no idea at present how white youth would benefit from the program if they were enrolled. But the effect size reported here is nonetheless quite striking. What makes this perhaps more remarkable still is that for logistical reasons, programming did not start until November 19, a week before Thanksgiving and thus lasted less than a full academic year.

It is noteworthy that our estimated gains in math test scores (0.65 of a control group SD) fit very comfortably alongside the non-experimental results reported by Fryer (2011) for the same high-dosage tutoring intervention, which in his Houston Public Schools study equaled 0.48 and 0.74 SD for 6th and 9th graders, respectively. Our findings are also consistent with the sizable gains in achievement test scores (0.60 SD) reported by Banerjee et al. (2007) from providing 3rd and 4th graders who are very far behind with remedial instruction that is better targeted to their level.

Nonetheless our results come from a small-scale pilot test and so these findings are necessarily not the last word on this subject; many questions remain. For example one key question is to understand more about the relative importance and effectiveness of the two separate components of our intervention. Our small sample size and the possibility of crossover and spillover effects across the two treatment arms complicates our ability to disentangle the relative value of the two components. And we have evaluated outcomes measured just during the program year; at present we know nothing about the degree to which these impacts will persist.

A general concern in education is whether interventions can succeed at large scale. The results reported here come from a pilot RCT at a single Chicago high school. We chose to work with this particular high school because of the outstanding school leadership team, which all else equal might make us think that our results may overstate the effects that we would see if the interventions were delivered in a broader set of schools where

the average quality of school leadership is somewhat lower. On the other hand the community in which this school is located is extremely disadvantaged, indeed one of the most distressed parts of the distressed south side of Chicago. Obviously there is no substitute for testing the intervention at scale, which is exactly what our research team is currently carrying out in multiple CPS high schools.

At the very least we are confident that the interventions reported on here *can* be delivered at large scale. The BAM program has been delivered to at least 700 students per year since 2009 in Chicago, including in some of the most distressed schools in the city, and has been previously evaluated at large scale and found to have impacts on behavioral outcomes, although not test scores (Heller et al., 2013). Learning more about the active ingredients in the curriculum and key characteristics of providers that facilitate success would obviously be of enormous value to facilitating even larger scale-up. The Match tutoring intervention is delivered to over 700 students per year across three schools in Boston, to 500-600 students in several public high schools in Lawrence, Massachusetts, and was delivered to 3,000 students across 13 schools in Houston (Fryer, 2011). A specific concern with the ability of providers to scale up "no excuses" schools has been whether there is sufficient supply of the right sort of provider; Match reports receiving 10 to 20 applications per opening for its work in Houston and Lawrence, and (for hiring on a shorter timetable) 5 applications per opening in Chicago.

Despite the remaining questions that this study alone cannot answer, it is important to recognize just how large these impacts are (on a per dollar of spending basis) compared to other interventions that have been tried with disadvantaged youth, or with younger children for that matter. Our best estimate for the cost per participant in our intervention is roughly $4,400, with a defensible range of $3,000 to $6,000.[18] The test score gain per dollar spent from this intervention are very large compared to previous interventions for disadvantaged

[18] The costs of the BAM CBT intervention is $1,900 per participant per academic year, which is a fairly reliable cost estimate that comes from Youth Guidance's experiences serving on average around 700 youth per year since the 2009-10 academic year. The costs of the academic intervention are somewhat harder to determine. Our research team delivered our best approximation of Match tutoring at probably inefficiently small scale, in the sense that we had "too much" supervisory capacity given the number of tutors and youth serviced. Our realized cost was about $4,000 per student, although we had fewer students participate than we had built capacity to serve – had we filled up each program slot, the cost would have been more like $2,800 per student. Whether to use the lower or higher figure depends partly on whether the underutilized tutor capacity wound up increasing the intensity of the intervention or was just idle. The cost per student for the actual Match tutoring intervention studied at scale by Fryer (2011) is reported to be $2,500 per student. The Fryer estimate plus the YG cost for BAM is $4,400 per youth per year. If we take three-quarters of that cost to account for our having delivered the intervention for just three-quarters of the year, the cost would be more like $3,330. As an upper bound we could add the YG full-year cost to our high-end full-year realized cost of tutoring for a total cost per participant of nearly $6,000.

youth; aside from Fryer's (2011) non-experimental study of the same academic intervention we examine here, we know of no intervention from a credible study that shows test score gains for this population *and* also reports on program costs.[19]

From the perspective of improving outcomes measured during adolescence, the size of the impacts per dollar spent we reported on here are large even in relation to some of the most successful early childhood interventions that have been studied. For example Figures 6 and 8 compare the estimated impacts per dollar spent on math test scores and overall course grades (GPA) in 10th grade from our intervention with those of Perry Preschool (Schweinhart et al., 2005), cash transfers from the Earned Income Tax Credit, or EITC (Dahl and Lochner, 2012), and class-size reduction in early elementary school grades (Krueger, 1999, 2003b; Schanzenbach, 2006); see Figures 6 and 8. These estimates focus on the present value of the cost per intervention calculated at 10th grade, assuming a 3% discount rate. Our intervention looks quite cost-effective in this comparison regardless of whether we use the low, middle or high end of our plausible range for program costs. While our intervention does not have statistically significant effects on reading scores, our estimate is also not very precise. The 95% confidence interval for our estimate (Figure 7) includes the point estimates for the effects from these other interventions.[20]

---

[19] While a few interventions have been shown to boost high school graduation rates for youth (see Krueger, 2003a; Guryan, 2004; Bloom, Muller-Ravett and Broadus, 2011; Murnane, 2013), few credible studies report statistically significant gains in standardized test scores for disadvantaged youth. One exception is Nomi and Allensworth's (2009) study of double-dose algebra in Chicago high schools, which found effect sizes on math scores of 0.26SD. Unfortunately nothing is reported about the cost of that policy, so we cannot compare the test score gain per dollar spent of that intervention to ours.

[20] The widely-cited Perry Preschool improves high school GPA by 0.42 SD, or 0.3 points on a 4-point scale, improves overall math scores at age 14 by 0.33 SD, improves reading scores at 14 by 0.34 SD, and language achievement by 0.63 SD, at a cost of $20,500 (in 2013 dollars) per child (Schweinhart et al., 2005, p. xvii, 62). We report the larger of the two "reading" scores in our figure (language arts) to be conservative. Their study does not report their regression-adjusted standard errors, so we calculate rough approximations of 95% CI's for the figure above assuming that impacts that have $p<.05$ have t-statistics=2.2, and impacts with $p<.1$ we assume to have t-statistics = 1.75. Obviously these are just approximations. Dahl and Lochner (2012) report that each $1,000 gain in contemporaneous family income boosts test scores for children ages 4-14 by 0.0359 SD in reading recognition (se=0.0195), 0.0613 SD in reading comprehension (se=0.0273), and 0.058 SD in math (se=0.0273). We report the larger of the two reading scores (comprehension) in our figure to be conservative. They find substantial decay in the effects of family income on achievement test scores so the impacts in our figure above would overstate (substantially) the gain in adolescent test scores that would come from increasing family income for children between the ages of 4-14. Smaller class sizes in grades K-3 in the Tennessee STAR experiment increased combined SAT/ACT test scores for blacks in high school, after accounting for treatment effects on test-taking rates, by 0.15 SD, at a per-child total cost over the whole intervention period of $19,645 in 2013 dollars (Schanzenbach, 2006). She does not report the standard error for this estimate (which is reported in the text of the paper) but the standard error for the grade 3 impacts (which is of about the same magnitude) is 0.03, which we use as an approximation in constructing the figure above. Since this is an overall test score (not subject-specific) we use these results for both the reading and math panels of our figure.

These comparisons highlight a tradeoff between intervening early on in childhood versus later during adolescence, which has not received enough attention in the current policy discussion. Much of the discussion has focused on the possibility of declining developmental plasticity as children age, including in the landmark National Academy of Sciences report *Neurons to Neighborhoods* (Shonkoff and Phillips, 2000), which has helped direct a great deal of attention to early childhood intervention. A growing body of recent research suggests a great deal of developmental plasticity during adolescence as well (for example Selemon, 2013, Steinberg, forthcoming), but for the sake of argument let us suppose for the moment that people really are more plastic and receptive to intervention during early childhood than during adolescence.

Given that most social-policy interventions show "fade out" of impacts, if the goal is to improve the *long-term* life outcomes of children growing up in disadvantaged circumstances, then there is a tradeoff between fade-out and developmental plasticity. Put differently, as a conceptual matter it is not obvious that early childhood is the optimal time to intervene even if there is declining developmental plasticity over the life course; precisely because the impact of interventions tend to fade out over time, there are some advantages to intervening temporally closer to when important outcomes are realized. It is also the case that many socially costly outcomes, particularly criminal behavior, are highly concentrated within the population. Another tradeoff between intervening early versus later comes from the fact that it is easier to target interventions on the highest-risk students during adolescence than early childhood, because we have more of a "track record" to use to identify participants who might benefit most. More research is still needed to better understand this tradeoff and the larger question of how to allocate intervention resources across the lifecycle in a way that generates the largest improvements in long-term life outcomes for disadvantaged children.

Our findings also highlight a systemic challenge for so many urban school districts – the need for a more intensive safety net to help students who fall behind as they progress through school and wind up experiencing a mismatch between what they need and what regular classrooms deliver. This mismatch is a problem that many previous interventions largely ignore, instead focusing on changing the quality of grade-level instruction in the classroom or the incentives of students to learn it. Efforts to address this mismatch on both the academic and

non-academic sides in our intervention show it is possible to generate very large gains in academic outcomes in a short period of time, even among students who can be many years behind grade level.

The key to making this intensive remediation affordable, particularly the high-dosage two-on-one individualized instruction, is the recognition that the tutoring method of instruction substantially changes the set of skills and experiences required to succeed as an instructor. We recognize that we cannot distinguish the effects of the Match academic intervention from the BAM non-academic intervention. But our results provide at least suggestive support for the idea that even first-year instructors with no formal teaching credentials or experience who are working for a modest stipend mostly as a public service can contribute to very large gains in student learning among a population (disadvantaged youth) for which there are so few previous success stories. There is, in short, the possibility of a tradeoff between the "teaching technology" used to deliver academic instruction and teacher "quality" the way so much of current education policy seems to define it.

Another factor that helps control the costs of incorporating this sort of intensive remediation into an urban school system's safety net to help students who fall behind is that the need for help might need to be only temporary. Our pilot experiment and Fryer's quasi-experimental study in Houston both suggest that students may be learning the equivalent of about three years' worth of math in a single year. In contrast students who are four to 10 years behind grade level, as unfortunately is not uncommon in distressed urban areas, have basically been getting little to nothing out of regular classroom instruction for years. If it is possible to achieve at large scale the results we report here, in which students learned the equivalent of three years of math per year, just a few years of this type of intervention could bring almost all students up to grade level – at which point they could begin to re-engage and benefit from the grade-level material taught in regular high school classrooms.

The large gains in academic outcomes for disadvantaged youth reported here stand against a backdrop of few prior success stories in improving academic outcomes, particularly achievement test scores, for similarly disadvantaged adolescents. The costs of even the bundled intervention, while not trivial, are not prohibitively high. The impacts per dollar spent are sizable compared to even the most successful early childhood programs. Perhaps the growing pessimism about academic interventions for low-income youth is premature, now that we may be diagnosing the key underlying problems.

REFERENCES

Allensworth, Elaine M. and John Q. Easton (2005) *The On-Track Indicator as a Predictor of High School Graduation*. Consortium on Chicago School Research at the University of Chicago.

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

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

Banerjee, Abhijit V., Shawn Cole, Esther Duflo and Leigh Linden (2007) "Remedying education: Evidence from two randomized experiments in India." *Quarterly Journal of Economics*. 122(3): 1235-64.

Barrow, Lisa, Amy Claessens, and Diane Whitmore Schanzenbach (2013) "The impact of Chicago's small high school initiative." Northwestern University, Institute for Policy Research Working Paper WP-13-20.

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

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

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli (2006) "Adaptive linear step-up procedures that control the false discovery rate." *Biometrika*. 93(3): 491-507.

Bloom, Benjamin S. (1984) "The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring." *Educational Researcher*. 13(6): 4-16.

Bloom, Dan, Sara Muller-Ravett, and Joseph Broadus (2011) *Staying on Course: Three-year results of the National Guard Youth ChalleNGe Evaluation*. New York, NY: MDRC.

Bloom, Howard S., "Accounting for No-shows in Experimental Evaluation Designs," *Evaluation Review*, 8 (1984), 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.

Carniero, Pedro and James J. Heckman (2003) "Human capital policy." In *Inequality in America: What Role for Human Capital Policies?* James J. Heckman and Alan B. Krueger. Cambridge, MA: MIT Press. pp. 77-240.

Cascio, Elizabeth U., & Staiger, Douglas O. (2012). *Knowledge, tests, and fadeout in educational interventions*. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 18038.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor (2009) "The academic achievement gap in grades 3 to 8." *The Review of Economics and Statistics*. 91(2): 398-419.

Coleman James S. et al. (1966) *Equality of Educational Opportunity*. Washington, DC: Government Printing Office.

Cullen, Julie B., Levitt, Steven D., Robertson, E., & Sadoff, S. (2013). What Can Be Done To Improve Struggling High Schools? *The Journal of Economic Perspectives*, *27*(2), 133-152.

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

Dahl, Gordon B. and Lance Lochner (2012) "The impact of family income on child achievement: Evidence from the Earned Income Tax Credit." *American Economic Review*. 102(5): 1927-56.

DeNavas-Walt, Carmen, Bernadette D. Proctor, and Jessica C. Smith (2011) *Income, Poverty, and Health Insurance Coverage in the United State: 2010 (Current Population Reoprts: Consumer Income. P60-239*). Washington, DC: US Department of Commerce, Economics and Statistics Bureau, US Census Bureau. http://www.census.gov/prod/2011pubs/p60-239.pdf

Dodge, K. A. (2003). Do social information-processing patterns mediate aggressive behavior? In B. B. Lahey, T. E. Moffitt, & A. Caspi (Eds.): *Causes of conduct disorder and juvenile delinquency*. New York: Guilford Press, 254-274.

Dodge, K. A., Bates, J. E., & Pettit, G. S. (1990). Mechanisms in the cycle of violence. *Science*, 250(4988), 1678-1683.

Dodge, K. A., Pettit, G. S., McClaskey, C. L., Brown, M. M., & Gottman, J. M. (1986). Social Competence in Children. *Monographs of the Society for Research in Child Development*, 51(2), i-85.

Duflo, Esther, Dupas, Pascaline & Kremer, Michael (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, 101(5), 1739-1774.

Duncan, Greg J. and Richard J. Murnane, Eds. (2011) *Whither Opportunity? Rising Inequality, Schools, and Chidlren's Life Chances*. New York: Russell Sage Foundation Press.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Iver, D. M. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist*, 48(2), 90-101.

Efron, R and R Tibshirani (1993) *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press.

Engel, M., Claessens, A., & Finch, M. A. (2012). Teaching students what they already know? The (Mis)Alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, 35(2), 157-178.

Evans, J. St. B.T. & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223-241.

Fryer, Roland G. (2010) "Racial inequality in the 21st Century: The declining significance of discrimination." Cambridge, MA: NBER Working Paper 16256.

Fryer, Roland G. (2011). Creating 'No Excuses' (Traditional) Public Schools: Preliminary Evidence from an Experiment in Houston, Cambridge, MA: NBER Working Paper No. 17494.

Glass, Gene V. (1976) "Primary, secondary and meta-analysis of research." *Educational Researcher*. 5(10): 3-8.

Guryan, Jonathan (2004) "Desegregation and black dropout rates." *American Economic Review*. 94(4): 914-43.

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, Sara B., Harold A. Pollack, Roseanna Ander, and Jens Ludwig (2013) "Preventing youth violence and dropout: A randomized field experiment." Cambridge, MA: NBER Working Paper 19014.

Hunt, D. E. (1975). Person-environment interaction: A challenge found wanting before it was tried. *Review of Educational Research*, 45(2): 209-230.

Jencks, Christopher and Meredith Phillips, Eds. (1998) *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

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

Kahneman, Daniel & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds). *Heuristics & Biases: The Psychology of Intuitive Judgment*. New York. Cambridge University Press: 49-81.

Keeley, Juliette (2011). Learning online in jail: A study of Cook County jail's high school diploma program. B.A. Thesis, University of Chicago.

Kling, Jeffrey R., Jeffrey Liebman, and Lawrence F. Katz "Experimental analysis of neighborhood effects," *Econometrica*, 75 (2007), 83-119.

Krueger, Alan B. (1999) "Experimental estimates of education production functions." *Quarterly Journal of Economics*. 114(2): 497-532.

Krueger, Alan B. (2003a) "Inequality, too much of a good thing." In *Inequality in America: What Role for Human Capital Policies?* James J. Heckman and Alan B. Krueger. Cambridge, MA: MIT Press. pp. 1-76.

Krueger, Alan B. (2003b) "Economic considerations and class size." *Economic Journal*. 113: 34-63.

Ladd, Helen F. (2012) "Education and poverty: Confronting the evidence." *Journal of Policy Analysis and Management*. 31(2): 203-227.

Lazear, Edward P. (2001) "Educational production." *Quarterly Journal of Economics*. 116(3): 777-803.

Levitt, Steven D. & Venkatesh, Sudhir (2000). An economic analysis of a drug-selling gang's finances. *Quarterly Journal of Economics*, 115(3), 755-789.

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.

Nomi, Takako and Elaine Allensworth (2009) "Double-dose Algebra as an alternative strategy to remediation: Effects on students' outcomes." *Journal of Research on Educational Effectiveness*. 2(2): 111-48.

Oreopoulos, Philip and Kjell G. Salvanes (2011) "Priceless: The nonpecuniary benefits of schooling." *Journal of Economic Perspectives*. 25(1): 159-84.

Papachristos, Andrew V. (2009). Murder by structure: Dominance relations and the social structure of gang homicide. *American Journal of Sociology*, 115(1), 74-128.

Reardon, Sean 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.

Schanzenbach, Diane Whitmore (2006). "What have researchers learned from Project STAR?" *Brookings Papers on Education Policy*, 205-228.

Schneider, W. & Shiffin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search and attention. *Psychological Review*, 84(1), 1-66.

Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Belfield, and Milagros Nores (2005) *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40*. Ypsilanti, Michigan: High/Scope Press.

Selemon, LD (2013) "A role for synaptic plasticity in the adolescent development of executive function." *Translational Psychiatry*.

Shiffrin, R. M. & Schneider, W. (1977). "Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory." *Psychological Review*, 84, 127–190.

Shonkoff, Jack P. and Deborah A. Phillips (2000) *From Neurons to Neighborhoods: The Science of Early Childhood Development*. Washington, DC: National Academy Press.

Steinberg, Laurence (forthcoming) *Our Last Best Chance: Why Adolescence Matters More than Ever*.

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

Westfall, P.H., and S.S. Young, *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Wiley-Interscience, 1993).

Figure 1. Frequency of BAM sessions attended, by treatment group

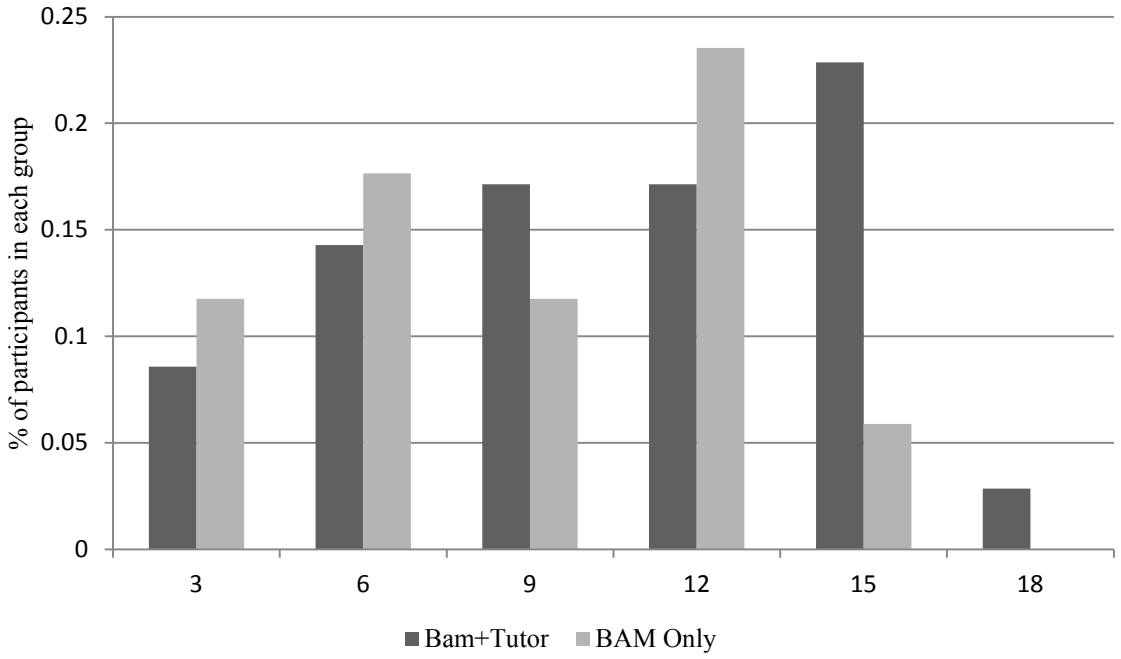

Figure 2. Frequency of tutoring sessions attended, youth assigned to BAM+Match group

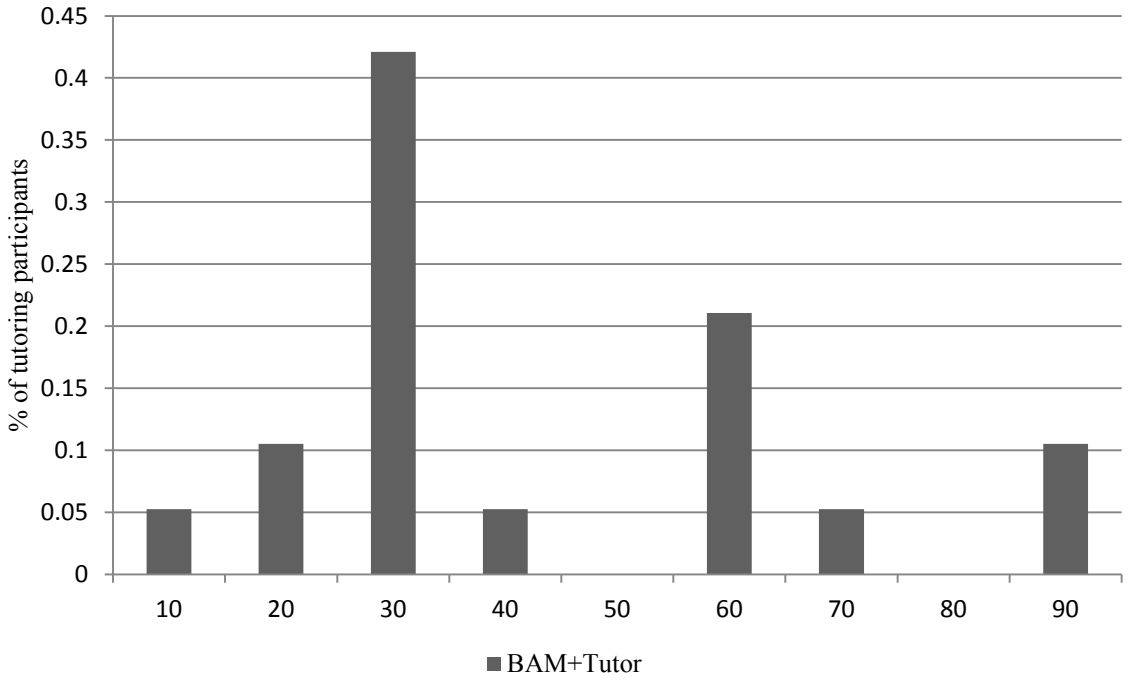

Figure 3. Intervention Boosts Math Test Scores by Sixty Percent of NAEP Black-White Test Score Gap

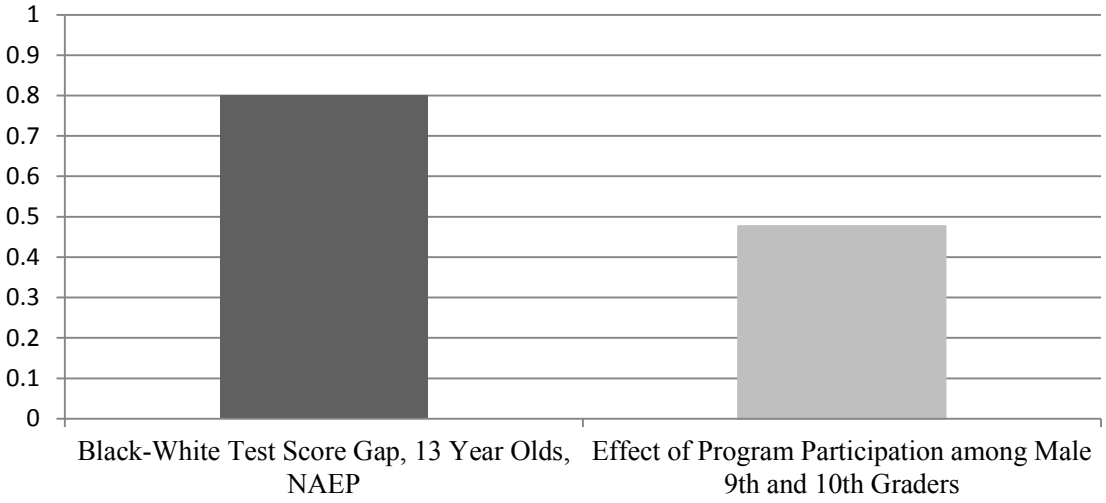

Figure 4: Effects of Program Participation on CPS "On Track" Indicator for HS Graduation

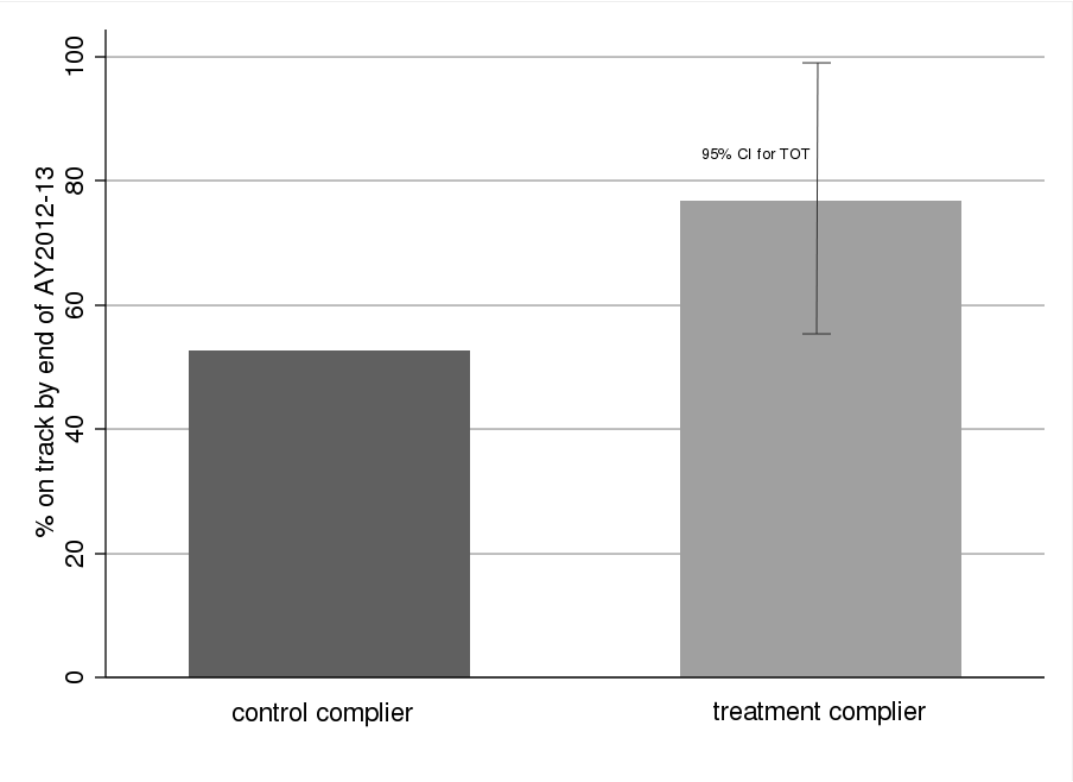

Figure 5. TOT Effect Sizes for BAM Only and BAM + Tutoring

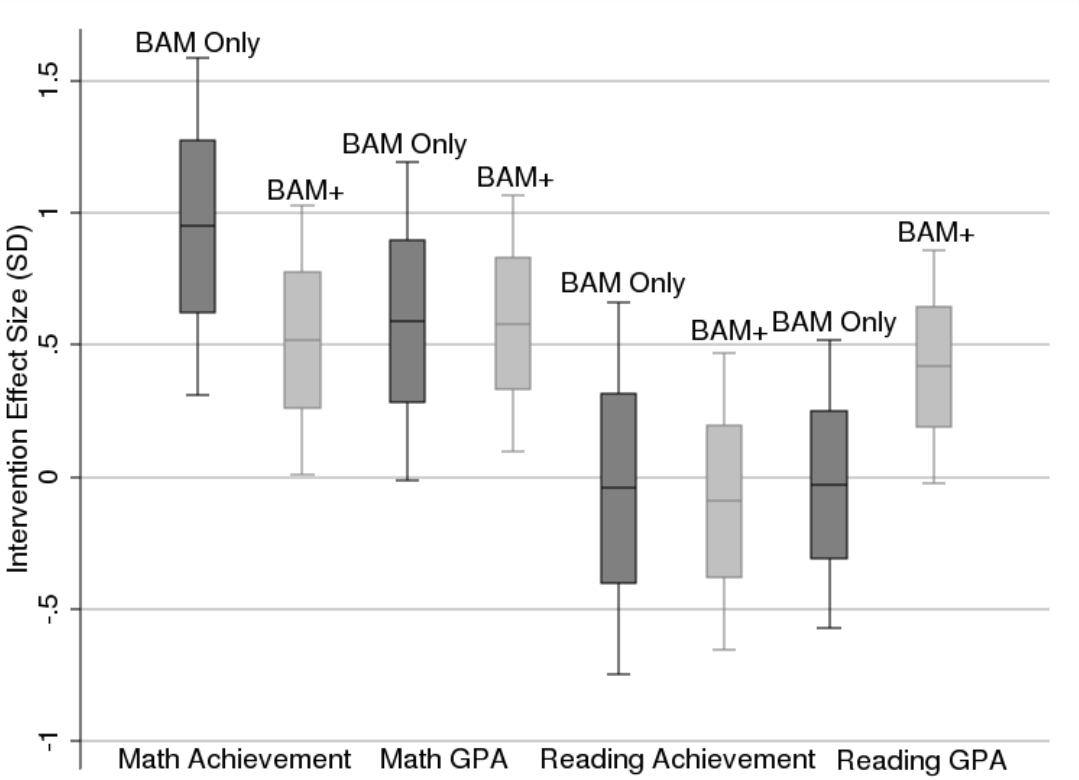

Figure 6. Impact on Math Test Scores During Adolescence, per $1,000 program cost (Z-Scores)

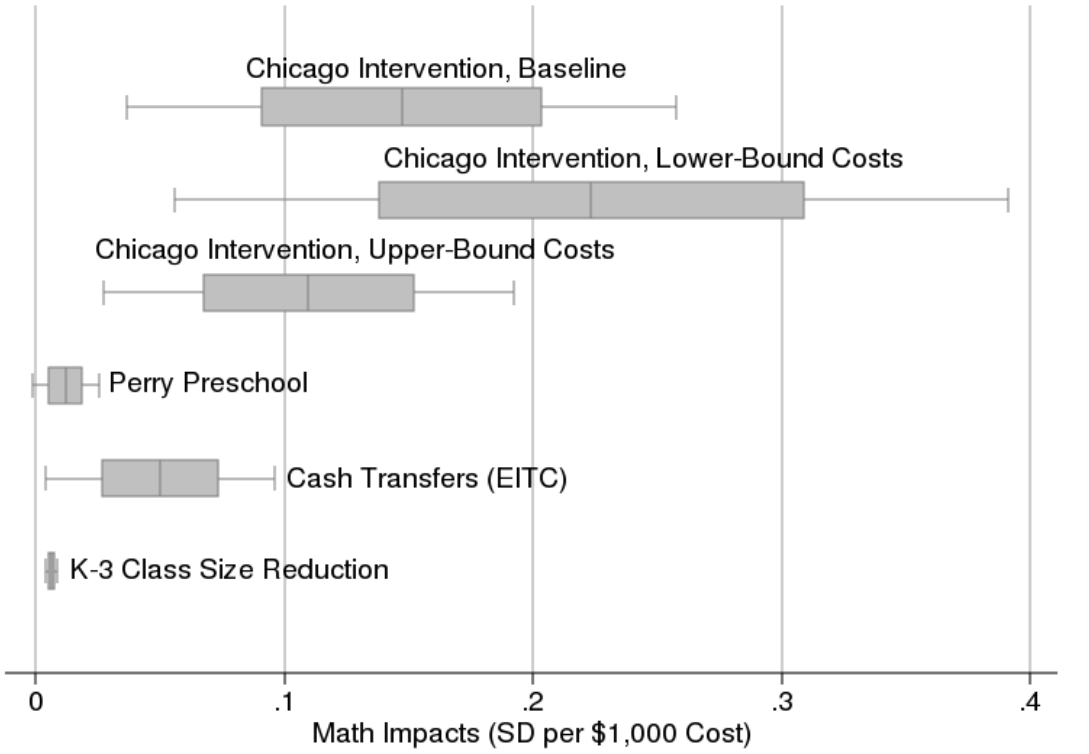

Figure7. Impact on Reading Test Scores During Adolescence, per $1,000 program cost (Z-Scores)

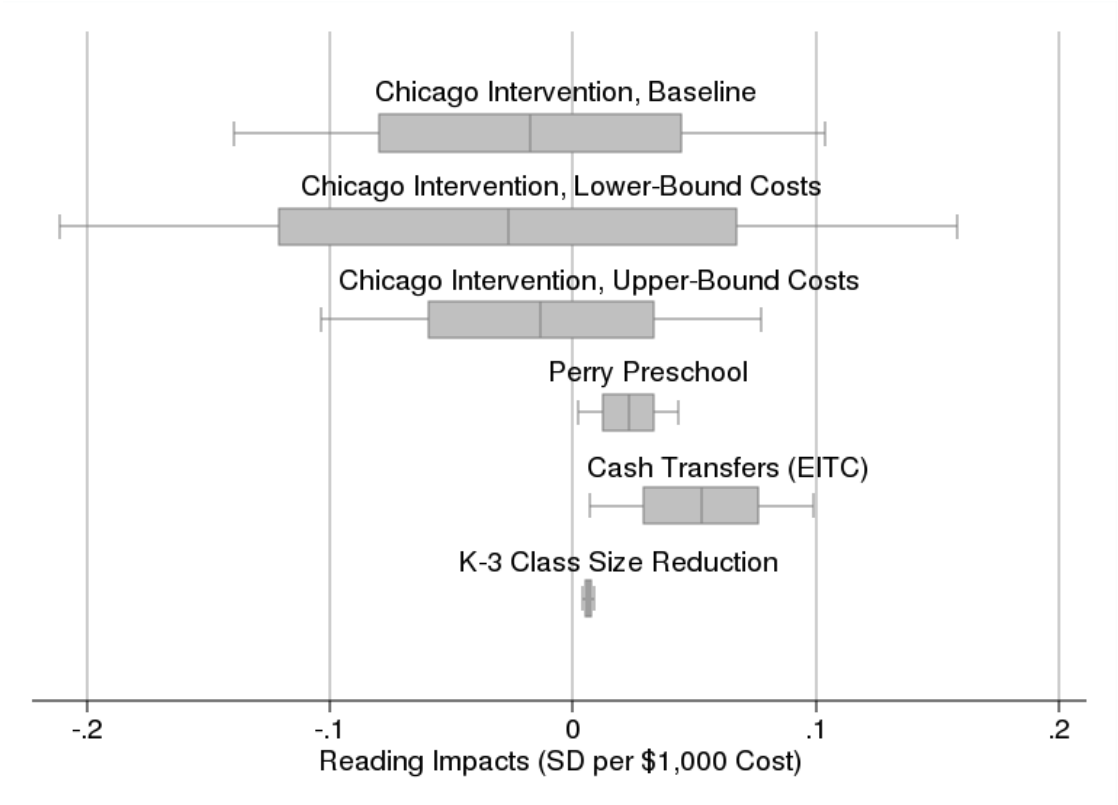

Figure 8. Impact on High School GPA per $1,000 of program costs

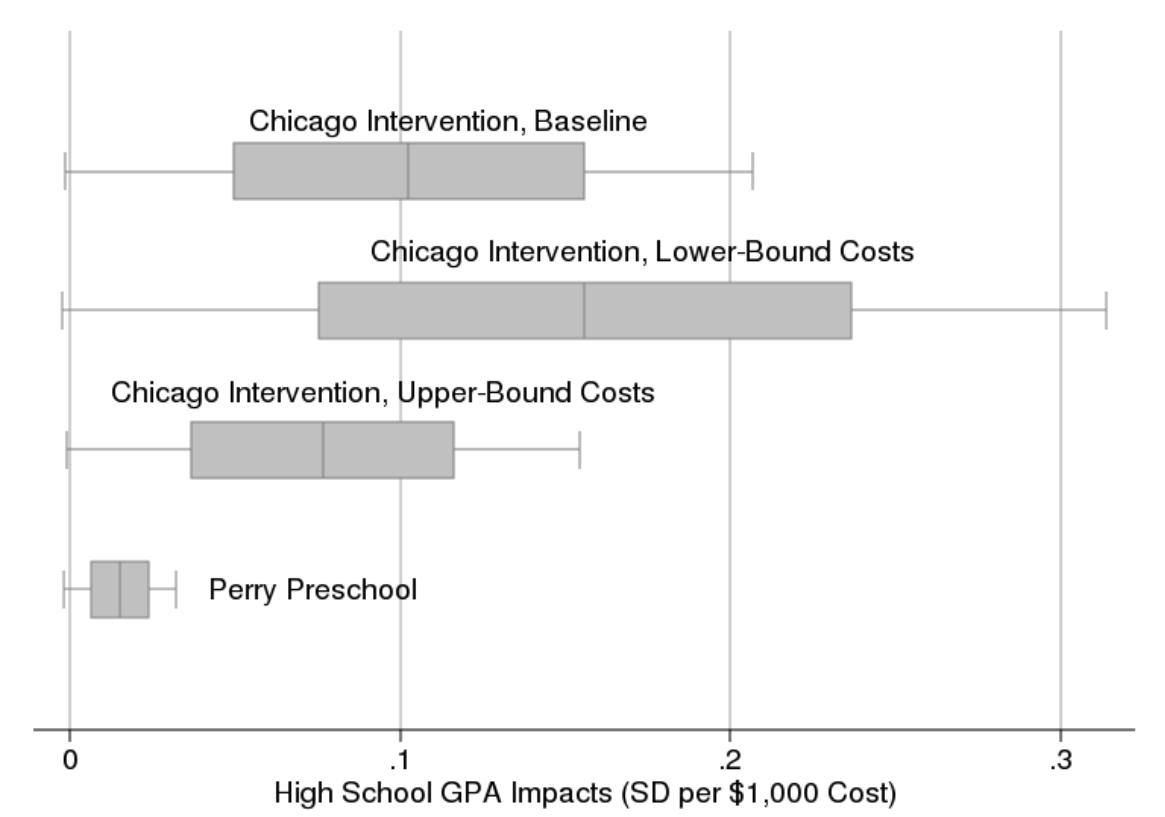

Table 1: Baseline characteristics by randomized groups

	All	Control	Assigned any treatment [1]	Assign BAM Only	Assign BAM + tutor
N Students	106	34	72	24	48
Age 14	0.27	0.35	0.24	0.25	0.23
Age 15	0.48	0.44	0.50	0.54	0.48
Age 16	0.25	0.21	0.26	0.21	0.29
Grade 10	0.56	0.56	0.56	0.46	0.60
Grade 9	0.44	0.44	0.44	0.54	0.40
Free lunch eligible	0.94	0.91	0.96	0.96	0.96
Reduced lunch eligible	0.05	0.09	0.03	0.04	0.02
Black	0.96	0.94	0.97	0.96	0.98
Hispanic	0.03	0.06	0.01	0.04	0.00
Other race	0.02	0.03	0.01	0.00	0.02
Learning Disability	0.26	0.26	0.26	0.25	0.27
GPA AY11-12	2.15	2.08	2.18	2.21	2.17
Non-math course GPA AY11-12	2.21	2.15	2.24	2.26	2.22
Math GPA AY11-12	1.86	1.75	1.92	1.92	1.92
All course failures AY11-12	1.73	1.88	1.66	2.00	1.49
Non-math course failures AY11-12	1.35	1.44	1.30	1.61	1.15
Math course failures AY11-12	0.38	0.44	0.36	0.39	0.34
Days Absent AY11-12	19.50	21.03	18.78	19.63	18.35
Discipline Incidents AY11-12	1.48	1.82	1.32	0.79 *	1.58
Out of School Suspension Days AY11-12	3.24	3.41	3.15	2.96	3.25
Fall Math Score, AY12-13, National Percentile	22.43	22.81	22.24	27.44	19.70
Fall Reading Score, AY12-13, National Percentile	26.22	27.41	25.64	30.72 +	23.16

+ $p<0.1$; * $p<0.05$; ** $p<0.01$

Students participated in the intervention during AY12-13. Student outcomes for AY11-12 occur the year before the program. Fall 2012 tests were administered before students were notified of their treatment or control status and services were provided

[1]Student assigned to either BAM only or BAM+tutoring

Table 2: Post-randomization mean outcomes by randomized groups

	All	Control	Assigned any treatment [1]	Assign BAM Only	Assign BAM + tutor
N Students	106	34	72	24	48
GPA AY12-13	1.67	1.48	1.75	1.68	1.79
Non-math course GPA AY12-13	1.69	1.55	1.75	1.66	1.80
Math GPA AY12-13	1.60	1.35	1.73 +	1.75	1.71
All course failures AY12-13	2.79	3.82	2.31 *	2.58	2.17
Non-math course failures AY12-13	2.28	3.18	1.86 +	2.04	1.77
Math course failures AY12-13	0.51	0.65	0.44	0.54	0.40
Days Absent AY12-13	37.10	44.03	33.74	33.39	33.92
Discipline Incidents AY12-13	1.48	1.56	1.44	1.54	1.40
Out of School Suspension Days AY12-13	1.19	1.81	0.89 +	0.68	1.00
Spring Math Score, AY12-13, National Percentile	32.37	25.29	35.35 +	45.39 *	30.72
Spring Reading Score, AY12-13, National Percentile	31.77	34.08	30.79	35.39	28.67
Has Spring 2013 Math Test	0.76	0.71	0.79	0.75	0.81
Has Fall 2012 Math Test	0.77	0.79	0.76	0.75	0.77
Has Both Fall 2012 and Spring 2013 Math Tests	0.66	0.65	0.67	0.67	0.67

+ $p<0.1$; * $p<0.05$; ** $p<0.01$

[1]Student assigned to either BAM only or BAM+tutoring

Students participated in the intervention during AY12-13. Symbols for significance denote that the pairwise comparison of treatment and control group differs significantly for that variable.

Table 3: Baseline means by randomized group and availability of spring 2013 post-tests in math

	control - has spring test	control - no spring test	Assigned any treatment [1] - has spring test	Assigned any treatment [1] - no spring test
N Students	24	10	57	15
Age 14	0.33	0.40	0.25	0.20
Age 15	0.54	0.20	0.53	0.40
Age 16	0.13	0.40	0.23	0.40
Grade 10	0.54	0.60	0.58	0.47
Grade 9	0.46	0.40	0.42	0.53
Free lunch eligible	0.92	0.90	0.95	1.00
Reduced lunch eligible	0.08	0.10	0.04	0.00
Learning Disability	0.25	0.30	0.26	0.27
GPA AY11-12	2.38	1.26	2.27	1.59
Non-math course GPA AY11-12	2.45	1.27	2.28	1.64
Math GPA AY11-12	2.15	0.90	2.00	0.96
All course failures AY11-12	0.83	4.80	1.40	3.20
Non-math course failures AY11-12	0.48	3.20	0.76	1.71
Math course failures AY11-12	0.13	0.80	0.20	0.50
Days Absent AY11-12	16.23	32.55	16.11	28.90
Discipline Incidents AY11-12	1.46	2.70	1.12	2.07
Out of School Suspension Days AY11-12	1.50	8.00	1.82	8.20
Has fall 2012 test results	0.92	0.50	0.84	0.47
Fall Math Score, AY12-13, National Percentile	24.73	14.40	23.94	10.57
Fall Reading Score, AY12-13, National Percentile	29.41	18.60	26.38	20.57

[1]Student assigned to either BAM only or BAM+tutoring

Students participated in the intervention during AY12-13. Student outcomes for AY11-12 occur the year before the program; fall 2012 tests were administered before students were notified of their randomization status or services provided.

Table 4: Estimated effects of program offer and participation on student learning outcomes and behavior during program year (AY2012-13)

	Control mean	Intent to treat (ITT)	Treatment on treated (TOT)	Control Complier Mean	Model p-value	Permutation test p-value	FWER-Adjusted p-value	FDR q-value
Outcome Domain: Math Achievement								
Math Achievement Test Scores Spring 2013 (Explore / Plan), N=81								
Z Score (Control Distribution)	0	0.510*	0.648**	-0.306	0.013	0.0163	0.03571	0.033
		[0.200]	[0.248]					
Z Score (National Distribution)	-1.059	0.375*	0.477**	-1.305	0.013	0.0162		
		[0.148]	[0.184]					
National Percentile Rank	25.292	11.930*	15.180**	18.602	0.015	0.0188		
		[4.774]	[5.860]					
Math GPA 2012-2013, N=105								
Math GPA 2012-2013 (1-4 point scale)	1.346	0.425*	0.583*	1.239	0.021	0.0261	0.0371	0.033
		[0.182]	[0.233]					
Math GPA Z Score (Control Distribution)	0	0.489*	0.670*	-0.123	0.021	0.0271		
		[0.209]	[0.267]					
Math Courses Failed 2012-2013, N=106								
Math Courses Failed 2012-2013	0.647	-0.301	-0.415+	0.684	0.117	0.1288	0.1173	0.041
		[0.190]	[0.238]					
Outcome Domain: Achievement in Other (Non-Math) Subjects								
Reading Achievement Test Scores Spring 2013 (Explore / Plan), N=81								
Z Score (Reading Test)	0	-0.061	-0.077	-0.076	0.798	0.804	0.799	0.441
		[0.236]	[0.273]					
Z Score (National Distribution)	-0.773	-0.037	-0.047	-0.837	0.775	0.781		
		[0.128]	[0.147]					
National Percentile Rank	34.083	-0.452	-0.575	31.684	0.918	0.919		
		[4.357]	[5.012]					
GPA in Non-Math Courses 2012-2013, N=106								
Non-Math GPA 2012-2013 (1-4 point scale)	1.547	0.207	0.285	1.582	0.204	0.217	0.366	0.257
		[0.162]	[0.205]					
Non-Math GPA Z Score (Control Distribution)	0	0.223	0.307	0.037	0.204	0.217		
		[0.174]	[0.221]					
Non-Math Courses Failed 2012-2013, N=106								
Non-Math Courses Failed 2012-2013	3.176	-1.454**	-2.004**	3.542	0.007	0.0094	0.019	0.022
		[0.531]	[0.701]					
Outcome Domain: Behavior								
Discipline Incidents AY12-13, N=106	1.559	0.121	0.167	1.025	0.723	0.735	0.725	0.438
		[0.341]	[0.438]					
Days Absent AY12-13, N=98	44.031	-10.272*	-12.919*	45.14	0.041	0.047	0.109	0.141
		[4.939]	[5.772]					
Out of School Suspensions Days AY12-13, N=98	1.813	-0.642	-0.808	1.596	0.203	0.219	0.336	0.255
		[0.501]	[0.580]					

+ p<0.1; * p<0.05; ** p<0.01

Covariates in all models: Indicator for age 14, indicator for age 15, indicator for grade 10, indicator for free lunch, indicator for learning disability, days absent AY11-12, days suspended AY11-12, discipline incidents AY11-12, GPA 2011-12, indicator for missing GPA AY11-12, fall AY12-13 math and reading Explore/Plan scores, indicator for missing Fall AY12-13 math and reading Explore/Plan scores

Standard errors reported in square brackets in the table. As treatment on treated (TOT) effects are essentially rescaled intent to treat (ITT) effects, we display ITT p-values in the interest of space and readibility. Intention to treat and treatment on the treated estimates are calculated as described in text, controlling for baseline covariates listed in Table 1.

Table 5: Sensitivity analyses, controlling for different combinations of baseline covariates

	Control mean	ITT - Full covariates	ITT controlling only for socio-demographics	ITT controlling only for prior schooling outcomes	ITT No covariates
Outcome Domain: Math Achievement					
Math Achievement Test Scores Spring 2013 (Explore / Plan), N=81					
Z Score (Control Distribution)	0	0.510*	0.438+	0.4	0.414+
		[0.200]	[0.228]	[0.241]	[0.245]
Z Score (National Distribution)	-1.059	0.375*	0.322+	0.271	0.297
		[0.148]	[0.167]	[0.176]	[0.180]
National Percentile Rank	25.292	11.930*	10.521+	9.607	10.059+
		[4.774]	[5.441]	[5.823]	[5.884]
Math GPA 2012-2013, N=105					
Math GPA 2012-2013 (1-4 point scale)	1.346	0.425*	0.439*	0.343+	0.380+
		[0.182]	[0.202]	[0.181]	[0.202]
Math GPA Z Score (Control Distribution)	0	0.489*	0.505*	0.394+	0.436+
		[0.209]	[0.233]	[0.208]	[0.232]
Math Courses Failed 2012-2013, N=106					
Math Courses Failed 2012-2013	0.647	-0.301	-0.269	-0.254	-0.203
		[0.190]	[0.199]	[0.183]	[0.202]
Outcome Domain: Achievement in Other (Non-Math) Subjects					
Reading Achievement Test Scores Spring 2013 (Explore / Plan), N=81					
Z Score (Reading Test)	0	-0.061	-0.172	-0.21	-0.189
		[0.236]	[0.255]	[0.267]	[0.262]
Z Score (National Distribution)	-0.773	-0.037	-0.097	-0.152	-0.119
		[0.128]	[0.138]	[0.148]	[0.146]
National Percentile Rank	34.083	-0.452	-2.507	-4.06	-3.294
		[4.357]	[4.775]	[5.018]	[4.987]
GPA in Non-Math Courses 2012-2013, N=106					
Non-Math GPA 2012-2013 (1-4 point scale)	1.547	0.207	0.255	0.139	0.205
		[0.162]	[0.188]	[0.161]	[0.188]
Non-Math GPA Z Score (Control Distribution)	0	0.223	0.275	0.15	0.221
		[0.174]	[0.203]	[0.174]	[0.203]
Non-Math Courses Failed 2012-2013, N=106					
Non-Math Courses Failed 2012-2013	3.176	-1.454**	-1.638**	-1.068+	-1.315*
		[0.531]	[0.576]	[0.561]	[0.637]
Outcome Domain: Behavior					
Discipline Incidents AY12-13, N=106	1.559	0.121	-0.109	0.106	-0.114
		[0.341]	[0.390]	[0.329]	[0.384]
Days Absent AY12-13, N=98	44.031	-10.272*	-12.591*	-8.039	-10.289+
		[4.939]	[5.440]	[5.090]	[5.552]
Out of School Suspensions Days AY12-13, N=98	1.813	-0.642	-0.940+	-0.655	-0.919+
		[0.501]	[0.523]	[0.473]	[0.507]

+ p<0.1; * p<0.05; ** p<0.01

Covariates in full covariate models: Indicator for age 14, indicator for age 15, indicator for grade 10, indicator for free lunch, indicator for learning disability, days absent AY11-12, days suspended AY11-12, discipline incidents AY11-12, GPA 2011-12, indicator for missing GPA AY11-12, fall AY12-13 math and reading Explore/Plan scores, indicator for missing Fall AY12-13 math and reading Explore/Plan scores

Covariates in socio-demographics models: Indicator for age 14, indicator for age 15, indicator for grade 10, indicator for free lunch, indicator for learning disability.

Covariates in prior schooling models: Days absent AY11-12, days suspended AY11-12, discipline incidents AY11-12, GPA 2011-12, indicator for missing GPA AY11-12

Table 6: Sensitivity analysis to multiple imputation and quantile regression with imputation

	Number of missing observations	Control mean	ITT	ITT (Imputed variables only)	ITT (Quantile Regression at Median, without imputed data)	ITT (Quantile Regression at Median, including imputed data)
Outcome Domain: Math Achievement						
Math Achievement Test Scores Spring 2013 (Explore / Plan), N=81						
Z Score (Control Distribution)	25	0	0.510*	0.374	0.422	0.393
			[0.200]	[0.276]	[0.289]	[0.299]
Z Score (National Distribution)	25	-1.059	0.375*	0.275	0.374+	0.28
			[0.148]	[0.209]	[0.199]	[0.225]
National Percentile Rank	25	25.292	11.930*	9.581	10.15	10.103
			[4.774]	[5.968]	[7.081]	[7.234]
Math GPA 2012-2013, N=105						
Math GPA 2012-2013 (1-4 point scale)	1	1.346	0.425*	0.453*	0.472*	0.548*
			[0.182]	[0.188]	[0.199]	[0.235]
Math GPA Z Score (Control Distribution)	1	0	0.489*	0.489*	0.543*	0.543*
			[0.209]	[0.209]	[0.228]	[0.229]
Outcome Domain: Achievement in Other (Non-Math) Subjects						
Reading Achievement Test Scores Spring 2013 (Explore / Plan), N=81						
Z Score (Reading Test)	25	0	-0.061	-0.08	-0.006	0.004
			[0.236]	[0.287]	[0.227]	[0.286]
Z Score (National Distribution)	25	-0.773	-0.037	-0.048	0.0032	-0.002
			[0.128]	[0.168]	[0.121]	[0.159]
National Percentile Rank	25	34.083	-0.452	-1.424	0.184	-1.29
			[4.357]	[5.73]	[4.734]	[5.83]
GPA in Non-Math Courses 2012-2013, N=106						
Non-Math GPA 2012-2013 (1-4 point scale)	0	1.547	0.207		0.21	
			[0.162]		[0.198]	
Non-Math GPA Z Score (Control Distribution)	0	0	0.223		0.226	
			[0.174]		[0.213]	
Outcome Domain: Behavior						
Discipline Incidents AY12-13, N=106	0	1.559	0.121		-0.24*	
			[0.341]		[0.334]	
Days Absent AY12-13, N=98	8	44.031	-10.272*	-9.93+	-9.57+	-8.64
			[4.939]	[5.054]	[4.931]	[5.382]
Out of School Suspensions Days AY12-13, N=98	8	1.813	-0.642	-0.77	-0.084	-0.86
			[0.501]	[0.519]	[0.462]	[0.434]

+ p<0.1; * p<0.05; ** p<0.01

ITT refers to the intent to treat estimates as calculated in Table 4 and are provided here for easy reference. ITT (imputed variables only) refers to models where missing data for outcome variables have been imputed using multiple imputation but where the model is otherwise the same. ITT (Quantile Regression at Median, without imputed data) estimates the ITT effect using quantile regression at the median of the data but without imputing missing data, and ITT (Quantile Regression at Median, including imputed data) estimates the ITT effect of treatment using quantile regression, but does so including imputed values for missing data.

Table 7: Results separately for each treatment arm

	Control mean	TOT (BAM Only)	TOT (BAM + Match)	p-value for test (TOT BAM-only = TOT BAM+Match)	p-value (TOT effects are equal, permutation test[1])
Outcome Domain: Math Achievement					
Math Achievement Test Scores Spring 2013 (Explore / Plan), N=81					
Z Score (Control Distribution)	0	0.949**	0.518*	0.141	0.186
		[0.325]	[0.259]		
Z Score (National Distribution)	-1.059	0.697**	0.382*	0.147	0.193
		[0.241]	[0.192]		
National Percentile Rank	25.292	22.057**	12.205*	0.155	0.202
		[7.696]	[6.135]		
Math GPA 2012-2013, N=105					
Math GPA 2012-2013 (1-4 point scale)	1.346	0.589+	0.581*	0.979	0.981
		[0.308]	[0.248]		
Math GPA Z Score (Control Distribution)	0	0.676+	0.667*	0.979	0.981
		[0.354]	[0.285]		
Math Courses Failed 2012-2013, N=106					
Math Courses Failed 2012-2013	0.647	-0.272	-0.475+	0.492	0.562
		[0.315]	[0.255]		
Outcome Domain: Achievement in Other (Non-Math) Subjects					
Reading Achievement Test Scores Spring 2013 (Explore / Plan), N=81					
Z Score (Reading Test)	0	-0.043	-0.092	0.88	0.893
		[0.360]	[0.287]		
Z Score (National Distribution)	-0.773	-0.03	-0.054	0.892	0.904
		[0.195]	[0.155]		
National Percentile Rank	34.083	-1.4	-0.218	0.8425	0.859
		[6.605]	[5.266]		
GPA in Non-Math Courses 2012-2013, N=106					
Non-Math GPA 2012-2013 (1-4 point scale)	1.547	-0.029	0.417+	0.087	0.115
		[0.278]	[0.225]		
Non-Math GPA Z Score (Control Distribution)	0	-0.031	0.450+	0.087	0.115
		[0.300]	[0.243]		
Non-Math Courses Failed 2012-2013, N=106					
Non-Math Courses Failed 2012-2013	3.176	-1.371	-2.271**	0.303	0.342
		[0.933]	[0.756]		
Outcome Domain: Behavior					
Discipline Incidents AY12-13, N=106	1.559	0.761	-0.082	0.124	0.154
		[0.585]	[0.474]		
Days Absent AY12-13, N=98	44.031	-13.441+	-12.705*	0.92	0.927
		[7.668]	[6.183]		
Out of School Suspensions Days AY12-13, N=98	1.813	-0.859	-0.787	0.921	0.928
		[0.770]	[0.621]		

+ p<0.1; * p<0.05; ** p<0.01

[1]100,000 replications

Table 8: Tests for treatment heterogeneity of ITT effect across different sub-groups defined by baseline characteristics

		Full covariates, no interactions	Interacting AY11-12 discipline incidents		Interacting AY11-12 out of school suspension days		Interacting AY12-13 discipline incidents		Interacting AY12-13 out of school suspension days	
	Control mean	Assigned BAM only	Assigned BAM only	Interaction	Assigned BAM only	Interaction	Assigned BAM only	Interaction	Assigned BAM only	Interaction
Outcome Domain: Math Achievement										
Math Achievement Test Scores Spring 2013 (Explore / Plan), N=81										
Z Score (Control Distribution)	0	0.611*	0.486+	0.176	0.423	0.214	0.458	0.196	0.515+	0.255
		[0.227]	[0.284]	[0.237]	[0.269]	[0.168]	[0.316]	[0.187]	[0.281]	[0.208]
Z Score (National Distribution)	-1.059	0.454*	0.394+	0.084	0.338+	0.132	0.349	0.139	0.384+	0.19
		[0.165]	[0.208]	[0.174]	[0.197]	[0.123]	[0.229]	[0.135]	[0.203]	[0.150]
National Percentile Rank	25.292	15.006**	9.828	7.322	9.186	6.640+	14.448*	2.575	14.973*	3.79
		[4.938]	[5.994]	[5.006]	[5.646]	[3.533]	[6.905]	[4.086]	[6.143]	[4.560]
Math GPA 2012-2013, N=105										
Math GPA 2012-2013 (1-4 point scale)	1.346	0.274	0.782**	-0.595**	0.608**	-0.152***	0.573*	-0.231+	0.596*	-0.364**
		[0.217]	[0.249]	[0.179]	[0.213]	[0.042]	[0.262]	[0.133]	[0.228]	[0.128]
Math GPA Z Score (Control Distribution)	0	0.315	0.898**	-0.683**	0.699**	-0.174***	0.659*	-0.265+	0.684*	-0.418**
		[0.250]	[0.286]	[0.206]	[0.245]	[0.048]	[0.301]	[0.153]	[0.262]	[0.148]
Math Courses Failed 2012-2013, N=106										
Math Courses Failed 2012-2013	0.647	-0.169	-0.520+	0.410+	-0.432	0.119*	-0.317	0.155	-0.318	0.227
		[0.250]	[0.309]	[0.223]	[0.265]	[0.052]	[0.338]	[0.171]	[0.308]	[0.173]
Outcome Domain: Achievement in Other (Non-Math) Subjects										
Reading Achievement Test Scores Spring 2013 (Explore / Plan), N=81										
Z Score (Reading Test)	0	-0.071	-0.033	-0.054	-0.171	0.114	-0.062	0.012	-0.029	-0.042
		[0.319]	[0.402]	[0.336]	[0.387]	[0.242]	[0.432]	[0.256]	[0.387]	[0.287]
Z Score (National Distribution)	-0.773	-0.039	-0.007	-0.045	-0.086	0.053	-0.039	0.009	-0.02	-0.021
		[0.172]	[0.217]	[0.181]	[0.209]	[0.131]	[0.233]	[0.138]	[0.208]	[0.155]
National Percentile Rank	34.083	-1.706	-0.941	-1.081	-3.631	2.196	-2.038	0.441	-0.902	-1.44
		[6.137]	[7.747]	[6.469]	[7.448]	[4.660]	[8.322]	[4.925]	[7.438]	[5.521]
GPA in Non-Math Courses 2012-2013, N=106										
Non-Math GPA 2012-2013 (1-4 point scale)	1.547	-0.117	0.332	-0.526**	0.218	-0.152***	0.164	-0.168	0.142	-0.213
		[0.199]	[0.230]	[0.166]	[0.190]	[0.038]	[0.257]	[0.130]	[0.234]	[0.132]
Non-Math GPA Z Score (Control Distribution)	0	-0.126	0.358	-0.567**	0.235	-0.164***	0.177	-0.181	0.153	-0.23
		[0.215]	[0.248]	[0.179]	[0.205]	[0.040]	[0.277]	[0.140]	[0.252]	[0.142]
Non-Math Courses Failed 2012-2013, N=106										
Non-Math Courses Failed 2012-2013	3.176	-0.573	-2.070*	1.754**	-1.591*	0.463**	-2.005*	0.911+	-1.459+	0.589
		[0.690]	[0.803]	[0.580]	[0.685]	[0.135]	[0.892]	[0.452]	[0.848]	[0.478]
Outcome Domain: Behavior										
Discipline Incidents AY12-13, N=106	1.559	0.485	-0.204	0.808+	0.271	0.097				
		[0.465]	[0.572]	[0.413]	[0.517]	[0.102]				
Days Absent AY12-13, N=98	44.031	-10.332	-18.173*	10.312	-15.077*	2.769	-15.490+	5.055	-10.099	0.107
		[6.739]	[8.310]	[6.609]	[7.175]	[1.657]	[8.307]	[4.210]	[7.800]	[4.395]
Out of School Suspensions Days AY12-13, N=98	1.813	-0.616	-0.902	0.375	-0.808	0.112				
		[0.641]	[0.811]	[0.645]	[0.702]	[0.162]				

+ p<0.1; * p<0.05; ** p<0.01

Table 8: Tests for treatment heterogeneity of ITT effect across different sub-groups defined by baseline characteristics

		Full covariates, no interactions	Interacting AY11-12 discipline incidents		Interacting AY11-12 out of school suspension days		Interacting AY12-13 discipline incidents		Interacting AY12-13 out of school suspension days	
	Control mean	Assigned BAM+tutor	Assigned BAM+tutor	Interaction	Assigned BAM+tutor	Interaction	Assigned BAM+tutor	Interaction	Assigned BAM+tutor	Interaction
Outcome Domain: Math Achievement										
Math Achievement Test Scores Spring 2013 (Explore / Plan), N=81										
Z Score (Control Distribution)	0	0.425+	0.552*	-0.093	0.516+	-0.055	0.558+	-0.056	0.417	0.049
		[0.226]	[0.265]	[0.100]	[0.263]	[0.079]	[0.317]	[0.166]	[0.277]	[0.092]
Z Score (National Distribution)	-1.059	0.316+	0.414*	-0.072	0.389*	-0.044	0.413+	-0.042	0.319	0.029
		[0.166]	[0.195]	[0.074]	[0.193]	[0.058]	[0.232]	[0.122]	[0.203]	[0.068]
National Percentile Rank	25.292	9.640+	13.081*	-2.521	12.336*	-1.614	16.024*	-3.861	11.218+	0.038
		[5.267]	[6.141]	[2.325]	[6.105]	[1.836]	[7.292]	[3.828]	[6.453]	[2.147]
Math GPA 2012-2013, N=105										
Math GPA 2012-2013 (1-4 point scale)	1.346	0.462*	0.164	0.173*	0.206	0.090*	0.389	-0.049	0.23	0.063
		[0.200]	[0.243]	[0.084]	[0.226]	[0.041]	[0.252]	[0.124]	[0.221]	[0.076]
Math GPA Z Score (Control Distribution)	0	0.530*	0.188	0.199*	0.237	0.104*	0.447	-0.056	0.265	0.073
		[0.229]	[0.279]	[0.097]	[0.259]	[0.047]	[0.290]	[0.143]	[0.254]	[0.087]
Math Courses Failed 2012-2013, N=106										
Math Courses Failed 2012-2013	0.647	-0.360+	-0.104	-0.146	-0.056	-0.096**	-0.29	-0.005	-0.21	-0.057
		[0.209]	[0.258]	[0.088]	[0.228]	[0.035]	[0.282]	[0.137]	[0.248]	[0.084]
Outcome Domain: Achievement in Other (Non-Math) Subjects										
Reading Achievement Test Scores Spring 2013 (Explore / Plan), N=81										
Z Score (Reading Test)	0	-0.043	0.092	-0.099	0.085	-0.077	-0.027	-0.017	-0.102	0.037
		[0.262]	[0.307]	[0.116]	[0.304]	[0.091]	[0.363]	[0.191]	[0.317]	[0.106]
Z Score (National Distribution)	-0.773	-0.026	0.045	-0.052	0.044	-0.042	-0.02	-0.009	-0.058	0.019
		[0.142]	[0.166]	[0.063]	[0.165]	[0.050]	[0.197]	[0.103]	[0.172]	[0.057]
National Percentile Rank	34.083	0.33	2.42	-1.532	2.825	-1.494	-0.129	0.18	-1.03	0.764
		[4.800]	[5.635]	[2.133]	[5.562]	[1.672]	[6.670]	[3.501]	[5.827]	[1.939]
GPA in Non-Math Courses 2012-2013, N=106										
Non-Math GPA 2012-2013 (1-4 point scale)	1.547	0.345*	0.103	0.138*	0.187	0.050+	0.262	-0.001	0.185	0.05
		[0.163]	[0.199]	[0.068]	[0.183]	[0.028]	[0.210]	[0.102]	[0.184]	[0.062]
Non-Math GPA Z Score (Control Distribution)	0	0.372*	0.111	0.149*	0.202	0.054+	0.283	-0.001	0.199	0.054
		[0.175]	[0.215]	[0.074]	[0.197]	[0.030]	[0.226]	[0.110]	[0.198]	[0.067]
Non-Math Courses Failed 2012-2013, N=106										
Non-Math Courses Failed 2012-2013	3.176	-1.782**	-0.609	-0.668**	-1.134+	-0.204*	-1.925*	0.275	-1.684*	0.085
		[0.563]	[0.669]	[0.229]	[0.628]	[0.096]	[0.768]	[0.373]	[0.680]	[0.230]
Outcome Domain: Behavior										
Discipline Incidents AY12-13, N=106	1.559	-0.008	0.257	-0.151	0.127	-0.042				
		[0.322]	[0.402]	[0.138]	[0.369]	[0.056]				
Days Absent AY12-13, N=98	44.031	-11.051+	-5.913	-3.208	-7.234	-1.25	-10.684	0.798	-10.266	0.432
		[5.556]	[6.816]	[2.490]	[6.230]	[0.944]	[7.561]	[3.669]	[6.674]	[2.263]
Out of School Suspensions Days AY12-13, N=98	1.813	-0.528	0.447	-0.609*	-0.312	-0.071				
		[0.590]	[0.702]	[0.256]	[0.669]	[0.101]				

+ p<0.1; * p<0.05; ** p<0.01

Appendix Table 1: all pre-randomization covariates

B / [SE] / (t) / <p>	Assign Any TRT	Assign BAM only	Assign Bam Tutor
Has Spring Achievment Test	0.155	0.111	0.169
	[0.161]	[0.259]	[0.191]
	-0.96	-0.43	-0.89
Age 14	-0.353+	-0.429	-0.408
	[0.207]	[0.359]	[0.259]
	(-1.71)	(-1.19)	(-1.57)
Age 15	-0.136	-0.191	-0.203
	[0.145]	[0.269]	[0.178]
	(-0.94)	(-0.71)	(-1.14)
Grade 10	-0.171	-0.2	-0.222
	[0.161]	[0.271]	[0.204]
	(-1.07)	(-0.74)	(-1.09)
Free luinch eligible	-0.373	*omitted*	-0.438
	[0.536]		[0.583]
	(-0.70)		(-0.75)
Reduced lunch eligible	-0.604	-0.144	-0.765
	[0.588]	[0.322]	[0.660]
	(-1.03)	(-0.45)	(-1.16)
Black	-0.275	0.172	-0.385
	[0.529]	[0.355]	[0.583]
	(-0.52)	-0.48	(-0.66)
Hispanic	-0.678	*omitted*	-0.999
	[0.602]		[0.710]
	(-1.13)		(-1.41)
Learning disability	0.035	0.075	-0.038
	[0.139]	[0.231]	[0.173]
	-0.25	-0.32	(-0.22)
GPA AY11-12	0.546	1.959	0.401
	[1.049]	[1.747]	[1.286]
	-0.52	-1.12	-0.31
Non-math course GPA AY11-12	-0.522	-1.623	-0.412
	[0.881]	[1.454]	[1.073]
	(-0.59)	(-1.12)	(-0.38)
Math GPA AY11-12	-0.064	-0.302	-0.043
	[0.199]	[0.335]	[0.258]
	(-0.32)	(-0.90)	(-0.17)
All course failures AY11-12	0.031	-0.119	0.046
	[0.107]	[0.210]	[0.128]
	-0.29	(-0.57)	-0.36
Non-math course failures AY11-12	-0.053	0.151	-0.08
	[0.130]	[0.262]	[0.160]
	(-0.41)	-0.57	(-0.50)
Days Absent AY11-12	-0.004	0	-0.004
	[0.005]	[0.007]	[0.006]
	(-0.79)	(-0.04)	(-0.73)
Discipline Incidents AY11-12	-0.048	-0.105+	-0.027
	[0.036]	[0.056]	[0.041]
	(-1.33)	(-1.86)	(-0.65)
Out of School Suspension Days AY11-12	0.015	0.021	0.01
	[0.016]	[0.025]	[0.019]
	-0.95	-0.83	-0.53
Fall Math Score, AY12-13, National Percentile	0	0.004	-0.001
	[0.003]	[0.007]	[0.004]
	-0.12	-0.54	(-0.32)
Fall Reading Score, AY12-13, National Percentile	-0.002	0	-0.004

	[0.005]	[0.009]	[0.006]
	(-0.41)	-0.01	(-0.68)
Missing fall tests	0.036	0.167	-0.039
	[0.168]	[0.275]	[0.206]
	-0.21	-0.61	(-0.19)
Missing academic data, AY11-12	-0.13	0.18	-0.024
	[0.499]	[0.802]	[0.704]
	(-0.26)	-0.22	(-0.03)
_cons	1.707+	0.398	2.002+
	[0.917]	[0.804]	[1.060]
	-1.86	-0.49	-1.89
F statistic	0.56	0.56	0.57
P-value	0.936	0.911	0.925
R squared	0.12	0.22	0.17
N	106	58	82

Appendix Table 1: regression covariates

B / [SE] / (t) / <p>	Assign Any TRT	Assign BAM only	Assign Bam Tutor
Has Spring Achievment Test	0.192	0.114	0.25
	[0.152]	[0.235]	[0.181]
	-1.26	-0.48	-1.38
Age 14	-0.354+	-0.569+	-0.351
	[0.196]	[0.326]	[0.241]
	(-1.80)	(-1.74)	(-1.46)
Age 15	-0.147	-0.274	-0.176
	[0.138]	[0.243]	[0.168]
	(-1.07)	(-1.13)	(-1.05)
Grade 10	-0.132	-0.281	-0.092
	[0.141]	[0.231]	[0.175]
	(-0.93)	(-1.22)	(-0.53)
Free lunch eligible	0.131	*0.086*	0.163
	[0.208]	[0.293]	[0.244]
	-0.63	-0.29	-0.67
Learning disbaility	0.003	0.057	-0.079
	[0.130]	[0.215]	[0.158]
	-0.03	-0.27	(-0.50)
GPA AY11-12	-0.014	0.016	-0.034
	[0.080]	[0.120]	[0.109]
	(-0.17)	-0.13	(-0.31)
Days Absent AY11-12	-0.003	*-0.002*	-0.005
	[0.004]	[0.006]	[0.005]
	(-0.81)	(-0.27)	(-0.91)
Out of School Suspension Days AY11-12	0.013	0.023	0.008
	[0.015]	[0.023]	[0.017]
	-0.89	-0.97	-0.48
Discipline Incidents AY11-12	-0.044	-0.098+	-0.023
	[0.034]	[0.053]	[0.039]
	(-1.28)	(-1.85)	(-0.60)
Fall Math Score, AY12-13, National Percentile	0	0.003	-0.002
	[0.003]	[0.006]	[0.004]
	-0.02	-0.54	(-0.41)
Fall Reading Score, AY12-13, National Percentile	-0.002	0	-0.005
	[0.004]	[0.009]	[0.005]
	(-0.56)	-0.05	(-0.99)
Missing fall tests	0.051	0.154	-0.042
	[0.159]	[0.254]	[0.190]
	-0.32	-0.61	(-0.22)
Missing academic data, AY11-12	-0.044	0.122	0.031
	[0.426]	[0.659]	[0.637]
	(-0.10)	-0.18	-0.05
_cons	0.806*	0.649	0.810+
	[0.372]	[0.597]	[0.449]
	-2.17	-1.09	-1.8
F statistic	0.62	0.68	0.57
p-value	0.841	0.779	0.876
R squared	0.09	0.18	0.11
N	106	58	82

	Assigned to any treatment and participated	Assigned to any treatment and did not participate	Assigned to BAM Only and participated	Assigned to BAM+Tutoring and participated	Assigned to BAM Only and did not participate	Assigned to BAM+Tutoring and did not participate
nstudents	52	20	17	35	7	13
Age 14	0.25	0.20	0.18	0.29	0.43	0.08
Age 15	0.52	0.45	0.59	0.49	0.43	0.46
Age 16	0.23	0.35	0.24	0.23	0.14	0.46
Grade 10	0.56	0.55	0.53	0.57	0.29	0.69
Grade 9	0.44	0.45	0.47	0.43	0.71	0.31
Free lunch eligible	0.96	1.00	0.94	0.97	1.00	1.00
Reduced lunch eligible	0.02	0.00	0.06	0.00	0.00	0.00
Black	0.02	0.00	0.00	0.03	0.00	0.00
Hispanic	0.94	1.00	0.94	0.94	1.00	1.00
Other race	0.04	0.00	0.06	0.03	0.00	0.00
Learning Disability	0.27	0.25	0.35	0.23	0.00	0.38
GPA AY11-12	2.25	2.01	2.19	2.28	2.28	1.87
Non-math course GPA AY11-12	2.29	2.09	2.24	2.32	2.31	1.97
Math GPA AY11-12	2.06	1.58	1.95	2.11	1.86	1.42
All course failures AY11-12	1.44	2.20	1.50	1.41	3.14	1.69
Non-math course failures AY11-12	1.20	1.55	1.25	1.18	2.43	1.08
Math course failures AY11-12	0.24	0.65	0.25	0.24	0.71	0.62
Days Absent AY11-12	17.28	22.68	16.21	17.80	27.93	19.85
Discipline Incidents AY11-12	1.12	1.85	0.59	1.37	1.29	2.15
Out of School Suspension Days AY11-12	2.35	5.25	1.82	2.60	5.71	5.00
Fall Math Score, AY12-13, National Percentile	21.86	23.58	25.07	20.14	39.33	18.33
Fall Reading Score, AY12-13, National Percentile	24.26	30.58	28.47	22.00	42.00	26.78
GPA AY12-13	1.86	1.47	1.97	1.81	0.97	1.73
Non-math course GPA AY12-13	1.87	1.46	1.95	1.83	0.95	1.73
Math GPA AY12-13	1.82	1.46	2.03	1.72	1.07	1.69
All course failures AY12-13	1.81	3.60	1.71	1.86	4.71	3.00
Non-math course failures AY12-13	1.54	2.70	1.41	1.60	3.57	2.23
Math course failures AY12-13	0.27	0.90	0.29	0.26	1.14	0.77
Days Absent AY12-13	32.22	39.39	30.18	33.21	44.30	36.67
Discipline Incidents AY12-13	1.19	2.10	1.06	1.26	2.71	1.77
Out of School Suspension Days AY12-13	0.79	1.29	0.24	1.06	2.20	0.78
Spring Math Score, AY12-13, National Percentile	33.78	41.91	42.36	30.03	56.00	33.86
Spring Reading Score, AY12-13, National Percentile	31.11	29.45	33.50	30.06	42.00	22.29
Has Spring 2013 Math Test	0.88	0.55	0.82	0.91	0.57	0.54
Has Fall 2012 Math Test	0.83	0.60	0.88	0.80	0.43	0.69
Has Both Fall 2012 and Spring 2013 Math Tests	0.75	0.45	0.76	0.74	0.43	0.46

Appendix 3: TOT p-values for main results

	Model p-value	Permutation test p-value
Outcome Domain: Math Achievement		
Math Achievement Test Scores Spring 2013 (Explore / Plan), N=81		
Z Score (Control Distribution)	0.009	0.0163
Z Score (National Distribution)	0.01	0.017
National Percentile Rank	0.01	0.0173
Math GPA 2012-2013, N=105		
Math GPA 2012-2013 (1-4 point scale)	0.012	0.019
Math GPA Z Score (Control Distribution)	0.012	0.019
Math Courses Failed 2012-2013, N=106		
Math Courses Failed 2012-2013	0.081	0.104
Outcome Domain: Achievement in Other (Non-Math) Subjects		
Reading Achievement Test Scores Spring 2013 (Explore / Plan), N=81		
Z Score (Reading Test)	0.777	0.797
Z Score (National Distribution)	0.752	0.775
National Percentile Rank	0.909	0.909
GPA in Non-Math Courses 2012-2013, N=106		
Non-Math GPA 2012-2013 (1-4 point scale)	0.165	0.196
Non-Math GPA Z Score (Control Distribution)	0.165	0.196
Non-Math Courses Failed 2012-2013, N=106		
Non-Math Courses Failed 2012-2013	0.004	0.007
Outcome Domain: Behavior		
Discipline Incidents AY12-13, N=106	0.703	0.727
Days Absent AY12-13, N=98	0.025	0.037
Out of School Suspensions Days AY12-13, N=98	0.164	0.2