
Institute for Policy Research
Northwestern University
Working Paper Series

WP-15-01

Not Too Late: Improving Academic Outcomes for Disadvantaged Youth

Philip J. Cook
Professor of Public Policy
Duke University

Kenneth Dodge
Professor of Public Policy
Duke University

George Farkas
Professor of Education
University of California, Irvine

Roland G. Fryer, Jr.
Professor of Economics
Harvard University

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

Jens Ludwig
Professor of Social Service Administration, Law, and Public Policy
University of Chicago

Susan Mayer
Professor of Public Policy
University of Chicago

Harold Pollack
Professor of Social Service Administration
University of Chicago

Laurence Steinberg
Professor of Psychology
Temple University

Corresponding author: Jonathan Guryan, j-guryan@northwestern.edu

Version: February 2015

DRAFT
Please do not quote or distribute without permission.

Acknowledgements:

This paper was made possible by the generous support of the Laura and John Arnold Foundation, as well as the city of Chicago, the Chicago Center for Youth Violence Prevention, the Chicago Public Schools, the Edna McConnell Clark Foundation, the Crown Family, the EquiTrust Life Insurance Company, the Lloyd A. Fry Foundation, the Illinois Criminal Justice Information Authority, the Joyce Foundation, JPAL – North America, the Reva and David Logan Foundation, the John D. and Catherine T. MacArthur Foundation, the Spencer Foundation, grant number 2012-JU-FX-0019 from the Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, U.S. Department of Justice, and award number 1P01HD076816 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health. We are also grateful for operating grants to the University of Chicago Crime Lab from the MacArthur and McCormick foundations. For vital assistance in making the intervention possible, we thank Chad Adams, Roseanna Ander, Barbara Byrd-Bennett, Valerie Chang, Akeshia Craven, Rukiya Curvey-Johnson, Gretchen Cusick, Aarti Dhupelia, Mayor Rahm Emanuel, Michael Goldstein, Craig Howard, Tim Jackson, Barbara Kelley, Ed Klunk, Timothy Knowles, Tim Lavery, Stig Leschly, Jonathan Lewin, Julia Quinn, Arnaldo Rivera, Janey Rountree, Alan Safran, Julia Stasch, Sara Stoelinga, Elizabeth Swanson, Robert Tracy, Karen Van Ausdal, and John Wolf, as well as the staffs of the Chicago Public Schools and Match Education. Thanks to Kylie Klein and Stacy Norris for their help in accessing the data we analyze here, to Amanda Norton for her valuable assistance, to Kelsey Reid, Michael Rosenbaum, Robert Webber, and David Welgus for their contributions to the analysis, and especially to Nathan Hess for his amazing leadership of the data analysis reported here. Thanks to Brian Jacob and Lawrence Katz for helpful comments. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice, National Institutes of Health or any other funder who has supported this work. Any errors are of course our own.

Abstract

There is growing concern that improving the academic skills of children in poverty is too difficult and costly once they reach adolescence, and so policymakers should instead focus either on vocationally oriented instruction or else on early childhood education. Yet this conclusion might be premature given that so few previous interventions have targeted a key barrier to school success: "mismatch" between what schools deliver and the needs of youth, particularly those far behind grade level. The researchers report on a randomized controlled trial of a school-based intervention that provides disadvantaged youth with intensive individualized academic instruction. The study sample consists of 2,718 male ninth and tenth graders in 12 public high schools on the south and west sides of Chicago, of whom 95 percent are either black or Hispanic and more than 90 percent are free- or reduced-price lunch eligible. Participation increased math achievement test scores by 0.19 to 0.31 standard deviations (SD), depending on how the researchers standardize, increased math grades by 0.50 SD, and reduced course failures in math by one-half in addition to reducing failures in non math courses. While some questions remain, these impacts on a per-dollar basis—with a cost per participant of around $3,800, or $2,500 if delivered at larger scale—are as large as those of almost any other educational intervention whose effectiveness has been rigorously studied.

I. INTRODUCTION

Improving the schooling outcomes of disadvantaged youth is a top policy priority in the U.S, one that is also central to addressing a wide range of social problems including poverty, health, and crime involvement, and to reducing inequality in life chances. Unfortunately, high school graduation rates in the U.S. have not changed much over the past 40 years (Heckman & LaFontaine, 2010; Murnane, 2013), despite substantial growth over time in the labor market returns to schooling (Goldin & Katz, 2010).[1] The differences in graduation rates between minority and white children remain large, with a black-white gap of 8 percentage points and a gap between Hispanics and whites equal to nearly 9 percentage points (Murnane, 2013). These differences are mirrored by disparities in achievement test scores as well. By age 13, the black-white gap in the National Assessment of Educational Progress (NAEP) equals 0.62 standard deviations (SD) in reading and 0.80SD in math.[2] The disparity in achievement test scores between rich and poor (the 90th vs. the 10th percentiles of the income distribution) seems to have increased substantially since 1940 (Reardon, 2011).

While there is widespread agreement about the importance of this problem, there remains great uncertainty about the best way to solve it. Part of the challenge is that there are far too few interventions that have been convincingly shown to improve outcomes for low-income students, particularly interventions that target adolescence – the time period when some of the most socially costly outcomes like high school dropout (and delinquency and teen fertility) are concentrated. For example, the What Works Clearinghouse gives only 2 of the 19 dropout-prevention programs it reviewed the top rating for strong effects; the Coalition for Evidence-Based Policy does not list a single program for addressing graduation rates in its "Top Tier."

This shortage of success stories has led to growing concerns about the value of efforts to improve academic outcomes for economically disadvantaged youth. For example Cullen, Levitt, Robertson and Sadoff (2013) argue that rather than focus on college-bound academics for disadvantaged teens, secondary schools

[1] As Murnane (2013) notes, there has been some increase in high school graduation rates for recent cohorts of youth, but no one knows why. He argues that "there are several hypotheses, but to date, very little evidence to explain the increase in high school graduation rates over the first decade of the twenty-first century."

[2] 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, 2014; Reardon, 2011).

should focus on technical or vocational education. Carneiro 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 by adolescence it is too late and too costly to improve the academic outcomes of children in poverty 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. We hypothesize that there are important mismatches between what many students (especially those from disadvantaged urban areas) need, and what many current education policies try (or are able) to provide. In general, the variance in achievement among students increases as they progress in school (Cascio & Staiger, 2012), a problem that is even more pronounced in urban areas where economic disadvantage among many students affects the rate at which they learn. Many students gradually fall behind grade level, which makes it more difficult to keep up with subsequent grade-level instruction, which then causes youth to fall yet further behind, further widening the variance in student achievement in large urban districts.

The standard way that schools are organized for instruction – a teacher teaching a group of students in a classroom – makes it difficult to individualize instruction and so may not be optimally designed for classrooms of students with widely varying skills, knowledge, and educational needs. Most education reforms focus on improving the quality with which grade-level material is taught, or the incentives students have to learn it. Yet such changes may have little effect on students who are far behind grade level – "saying it louder" won't help. Despite the $590 billion the U.S. spends each year on public K-12 schooling,[3] most urban school systems lack adequate safety nets to intensively help those who have fallen behind – this remains a key systemic challenge.

In this paper we report on the results of a randomized controlled trial (RCT) that assigned 2,718 9th and 10th grade male students within 12 Chicago Public Schools (CPS) high schools to receive either intensive individualized instruction or to a control group. The intervention, which was developed and is delivered by Match Education of Boston, is intensive, individualized two-on-one math tutoring provided for one hour per

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

day each and every day, during the school day. This type of tutoring is also one of the elements of the package of five "no excuses" charter school reforms Fryer (2014) introduced into a set of Houston public schools with encouraging results. Treatment-group students enroll in Match tutoring as a for-credit class, which replaces either a second period of math ("double dose algebra") that many 9th grade control students get, or else an elective course. Both treatment and control groups have access to status quo CPS services like Title I-funded after-school tutoring, but that is typically not very intensive.

For decades education researchers have thought that small-group tutoring generates "the best learning conditions we can devise," but 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). The key insight of Match is to solve 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 City Year or AmeriCorps).

The contribution of our paper is to present evidence from a large-scale RCT that it is possible to substantially and cost-effectively improve academic outcomes for disadvantaged children *even once they reach adolescence*.[4] The estimated effects of treatment on the treated (TOT) for math achievement are on the order of 0.19 to 0.31 standard deviations, depending on the test and norming that we use. These impacts are measured by two separate broad-based tests of math achievement: the ACT Inc.'s EXPLORE and PLAN tests, which CPS administers to 9th and 10th graders, respectively, as well as in-person achievement tests designed for the U.S.

[4] Fryer (2014) randomly assigns 16 elementary schools in Houston to receive a package of no-excuses charter school reforms, such as changes in the school culture around having high expectations for students, use of data-driven instruction, additional learning time (longer school day), and changes in school leadership. Schools also received high-dosage Match tutoring (3-on-1 in this study) but because of budget constraints this was only delivered to 4th graders in the elementary schools. Table V of his paper shows that the gain in math scores for 4th graders who received tutoring plus the other no-excuses elements was 0.22 SD compared to 0.17 SD for 5th graders who received all of the other no-excuses elements but not tutoring (the p-value for the comparison was 0.38). His earlier quasi-experimental analysis of middle and secondary schools, which were not randomly assigned, suggests a larger difference between grades (6th and 9th) that did receive tutoring plus the other elements (0.61) versus just the other elements (7th and 10th grade, with effect size of 0.21). Our research team also found very encouraging results from a single-school pilot study in Chicago during the 2012-13 academic year in which students received a combination of Match tutoring and Youth Guidance's Becoming a Man (B.A.M.) non-academic intervention (Cook et al., 2014).

Department of Education's NELS:88 study that our survey subcontractor (the Institute for Social Research at the University of Michigan) administered to a randomly selected sub-sample of our larger study population. The fact that we see similarly sized impacts on two different tests is one indication that the results we see on CPS tests are not the result of a narrow "teaching to the test" by the Match tutors. A similar conclusion is suggested by the fact that we see very large improvements in math grades (0.50 SD) and sizable reductions in course failures both in math (one-half the control complier mean, or CCM) and also in non-math courses. There may also be a reduction in violent-crime arrests from participation in this tutoring. The point estimate implies a 60% reduction in arrests compared to the CCM, although the result is somewhat imprecisely estimated and so even though it is large it is only marginally significant.

Our findings 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 – male 9th and 10th grade students living in very economically disadvantaged circumstances, including some of the most distressed and dangerous communities in Chicago (or anywhere in the U.S.). Of the youth in our study sample, over 90% are eligible for free or reduced price lunch and almost all come from very racially and economically segregated neighborhoods. The year before our intervention (that is, during the 2012-13 academic year), the average youth in our sample had a GPA of about 2.1 on a 4-point scale and had missed about a month of school. Around one in five had been arrested prior to the start of the study.

Because this intervention involves partnering youth with pro-social adults in a very small-group setting, in principle one candidate mechanism through which youth outcomes might improve is through mentoring or the development of what Coleman (1988) called "social capital." Yet we find no evidence in our survey data that program participants are more likely than controls to report feeling connections to or supported by pro-social adults. Nor do we find evidence that the program increases exposure to pro-social peers. As best we can tell, the academic component of the intervention itself seems to be an important element in explaining the impacts that we observe here. Among the only mechanism measures that we see changing as a result of program participation are indications of how confident and positive youth are about math and getting good grades.

There are several important questions that remain about our results, including whether or how these results will persist over time. Yet our benchmark estimate for the cost of the Match intervention is on the order of $2,500 per student carried out at large scale in a district, and about $3,800 per student at the scale of our current Chicago study, suggesting that this strategy could be feasible at large scale. The impacts per dollar spent appear to be at least as large as almost any other educational intervention that has been rigorously tested.

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

II. THEORY

Our study is motivated by the hypothesis that the large variance we see in student achievement in public school systems – particularly urban school districts – creates a "mismatch" between the sorts of supports that many 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.

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 & Staiger, 2012). The result is great variability in academic levels and needs by middle or high school, which are particularly 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.[5] 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), 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.[6] 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 in *both* the top *and* bottom halves of the achievement distribution. This experiment suggests that for initially low-performing students the benefits in tracked schools from reducing academic mismatch (from better-targeted instruction) are not only important, but also large enough to outweigh any adverse peer effects from being in tracked schools surrounded by lower-achieving classmates.

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, and in particular to individualize 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

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

[6] 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. 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.

the drawback of being inflexible, and possibly reducing 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 carried out by Banerjee et al. (2007), which found that assigning third and fourth graders in India 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 0.60 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 both strategies were short-lived.

III. THE INTERVENTION

We selected Match Education's tutoring model for this study partly 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. Another important motivation was the combination of the low cost and high dosage of the Match tutoring model. A final motivation for selecting the Match intervention was Fryer's (2014) encouraging experimental findings for elementary school students in Houston and quasi-experimental findings for middle and high school students.

Bloom (1984) summarizes a series of RCTs with elementary and middle school students in which the students were taught new subjects in which they would have had little prior background (cartography and probability). He finds that students assigned to receive one-on-one or small-group (not more than three-on-one) tutoring earned 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 generated large increases in time-on-task (90+% versus 65%) and improved student attitudes and interest. 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 ensured that all students received this attention – including those

students who were 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 put it, is to identify lower-cost instructional alternatives that can be as effective as tutoring.

One major innovation of the Match model, and another 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. Compared to regular classroom instruction, two-on-one tutoring greatly simplifies the adult's instructional task. 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. This enables Match to expand their recruitment pool and focus on people who are talented, have strong math skills, and are willing to devote a year to public service (such as new college graduates, retirees or career-switchers), but who do not necessarily have extensive prior training or experience as teachers. As with other public service programs like City Year, the tutors are willing to work for relatively low wages ($17,000 plus benefits for the nine-month academic year). A school in which classroom teachers cycled in and out after just one year would be quite challenging because of the important on-the-job learning that occurs during the first several years of teaching, for example, learning about things like classroom management. But that on-the-job learning is plausibly less important for two-on-one tutoring. The Match Education 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). This makes the incredibly high dosage of the Match tutoring model feasible from a cost perspective.

Match Education of Boston delivered this tutoring intervention in Chicago starting in the 2013-14 academic year in 12 CPS high schools, under sub-contract with our research team.[7] During the school day, students were assigned to participate in a 55-minute-long tutoring session as part of their regular class schedule, every day. In the CPS system, that is up to 165 contact hours per year. Each tutor was assigned to work with two students at a time during each session. The focus of the first half hour of each tutoring session was on remediating students' skill deficits, for which Match has developed its own skill-building curriculum. The second half of each session was tied to what youth learn in their classrooms.[8] Match used frequent internal formative assessments of student progress to individualize instruction.[9] Tutors taught six periods a day and each school was overseen by a site director. Site directors handled behavioral issues in the tutoring room, communication with school staff, and offered daily feedback and professional development.[10]

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). Each tutor participated in 100 hours of training, receiving daily feedback and an opportunity to exit the program if Match staff or the tutor saw lack of fit. Of the 53 Match tutors hired to serve in our study schools in the 2013-14 school year, about half were from minority race / ethnic groups (18 were African American; 8 were

[7] Guidance about how to incorporate the intervention into the CPS system came from a small-scale pilot study our team carried out the previous academic year (2012-13), which involved delivering our own version of the tutoring model in one high school. Details are reported in Cook et al. (2014).

[8] For the Fryer study, the Match curriculum was reverse-engineered from Texas state standards by Fryer's EdLabs. For this project, an Illinois-certified teacher was hired to develop curricula based on Illinois state standards.

[9] These include *daily tickets to leave* (1- to 3-question mini-assessments of the day's lesson, which allows the tutor to revise the next day's lesson based on the prior day's learning); *pre-tests and post-tests* (Match divides the year into 7 to 10 "course units," each with a pre- and post-test, which show tutors how much review time to allow for the first 2-3 weeks of the next unit); *quarterly proficiency assessments* (Match has used for nine years an 80-question assessment of basic math skills administered at baseline and up to 3 other times in the year; tutors target specific areas the student has not yet mastered for the next quarter, until the student scores at least 90% on the test); and *site-specific norm-referenced tests* (which will involve interim assessments of the tests CPS uses).

[10] Each site director has some combination of experience including math teaching / tutoring, mentoring, program direction, nonprofit management, public speaking, and training of adults, and is fully trained in the Match model. Tutors complete a daily report to the site director; here they note each student's progress and convey any issues.

Hispanic). Additionally, 19 tutors spoke fluent Spanish, and every school with high proportion of Spanish speakers had multiple bilingual tutors.

We focused on math skills because that is the focus of Fryer's study, because failure to complete required core math classes is one of the key drivers of high school dropout in Chicago,[11] and because of growing research about the importance of math specifically for short- and medium-term success in school, and also for long-term life outcomes like employment and earnings (Duncan et al., 2007). We focused on male youth partly because their graduation rates and test scores lag behind those of girls.

The control group in our study sample was eligible for all the status quo supports currently offered to students in the 12 CPS high schools in our study. These services include No Child Left Behind (NCLB) funded tutoring, which is of much lower dosage (and without the same structure, curriculum, or supervision) than Match tutoring. We estimate ~25% of control students in our schools received this Supplemental Educational Services (SES) tutoring, which involves 21 hours of writing tutoring per year and 20 hours of math (i.e., a bit over one-half hour *per week* of math tutoring, compared to one hour *per day* with Match); previous non-experimental studies of SES tutoring in Chicago find little detectable effect on math scores.[12] While these schools include a variety of other programs (and both treatment and control groups are eligible for these other programs in all schools) none of them focus on academic skill development in the same way Match does.[13]

IV. DATA

To measure baseline characteristics, program participation, and outcomes, we rely on three main sources of data: administrative longitudinal student-level records from the Chicago Public Schools (CPS), arrest records from the Chicago Police Department (CPD), although in this version of the paper we only have those arrest data for an abbreviated follow-up period (we will update in a subsequent version), and provider records.

[11] http://www.nytimes.com/2012/07/29/opinion/sunday/is-algebra-necessary.html?pagewanted=all&_r=0

[12] The most recent evaluation of SES in Chicago is for 2005-6 (http://sesiq2.wceruw.org/documents/chicago_ses.pdf). Statewide data suggest only around 14 percent of all students (about 60% of students who applied for SES services) receive SES services (http://www.quickanded.com/wordpress/wp-content/uploads/2011/11/SESChart-31.pdf).

[13] Eight of the schools also included GEAR UP services, which focus on preparing for college (essay writing and ACT prep) rather than basic skill development. Four of the schools included Youth Guidance's Project Prepare, another college-readiness program (see http://www.youth-guidance.org/our-programs/project-prepare/), four of the schools provided Peace Circles to help youth resolve disputes, 11 of the schools included the Mikva Challenge, which focuses on civic education (http://www.mikvachallenge.org/), and one school included buildOn, which focuses on involving youth in community service (http://www.buildon.org/afterschool-service/chicago/).

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 schools, by construction we have CPS student ID numbers for everyone we randomly assigned. From CPS we obtained student-level school records for the two academic years preceding the intervention (2011-12 and 2012-13) as well as the academic year following random assignment, 2013-14 (the first intervention year). 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; race / ethnicity; eligibility for free and reduced price lunch; course grades in each subject; enrollment status; absences; and disciplinary actions and suspensions.

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).[14] 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).[15]

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 report impact estimates using the test score results scored in three different ways. First, we show test score results using the EXPLORE and PLAN scale scores normalized to our control group's distribution, that is, subtracting off the control mean from each

[14] 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

[15] 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

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.

One subtlety with interpreting these percentile rank results is that CPS administers the 9th grade EXPLORE and 10th grade PLAN tests in the spring, so that CPS can measure year-to-year growth in test scores through the 11th grade administration of the ACT test, while the percentile scores for at least the 9th grade EXPLORE tests are normed to the national distribution of scale scores from students who took the EXPLORE in the fall, not the spring. So the CPS 9th graders have almost a full additional academic year more of schooling than the norming population to whom they're being compared.[16] This issue should not affect our estimates for the effect of treatment-control differences from Match tutoring. Furthermore, though the level of the percentile score is hard to interpret, the difference in the percentile scores between the treatment and control group is probably a reasonable approximation of the size of the treatment effects in percentile points.

A different issue with this outcome measure is that we are missing post-randomization (spring 2014) CPS test score results for 799 students out of the full 2,718 sample size (29 percent). It is reassuring for purposes of internal validity that the share of students missing spring 2014 CPS test scores is similar across the treatment and control groups. But there are two other issues with the missing-ness in the CPS test score data that will affect the size of the TOT effects we see. One is that the students for whom we have valid post-randomization CPS test scores have a higher-take-up rate in Match compared to the rest of the sample (51 percent versus 15 percent), which will lead the ratio of the TOT to the ITT effect to be somewhat smaller for the CPS test scores compared to our other CPS outcomes. The other issue has to do with the external validity of our

[16] The percentiles we have for the PLAN test (10th grade) are based on a spring norming population. (ACT has both a fall and spring scale score to percentile table for PLAN, but only a fall one for EXPLORE.)

CPS test results, since as we show below (Table 3), the youth for whom we are missing spring 2014 CPS tests have much lower GPAs compared to students for whom we have valid spring CPS test results.

Our second source of administrative government data for this project is Chicago Police Department arrest records for the intervention year through July 1, 2014 (the latest arrest data we have access to at the time of writing this version of the paper). Because criminal behavior is disproportionately concentrated over the summer months, we expect the present draft's analysis of these arrest records to potentially under-state much of the intervention's impacts on criminal or delinquent behavior. We are working to update these CPD arrest records and will incorporate those results into the next draft of the paper.

Our final data source comes from in-person surveys carried out for our research team under sub-contract by the Institute for Social Research (ISR) at the University of Michigan. For budget reasons, we selected a random sub-sample of 881 youth out of our larger analysis sample for the ISR survey sampling frame. ISR used two-phase sampling in which after interviewing 70% of the survey sample frame in the first phase, they selected a random sub-sample of youth in the second stage for more intensive follow-up. All of our analyses using these data employ sampling weights that account for this two-phase sampling design. ISR completed surveys with 658 youth, for an effective response rate of 88.1%. Most of the surveys were completed in May through June 2014 (the end of the first intervention year), although some were also completed in fall 2014. Month of survey completion is fairly balanced across randomized groups, as is the overall survey response rate.

This in-person data collection included a number of questions related to candidate mechanisms of action through which Match might affect schooling outcomes and other behavior, discussed below, as well as our own broad-based math achievement tests. We believed these math assessments would be an important complement to the CPS-administered EXPLORE and PLAN tests partly because of the possibility of floor effects on the CPS tests (students who are 7 or 10 years behind grade level in their math abilities could show large gains in math skills from the intervention but not show any gains on the EXPLORE or PLAN tests if most of those tests' material is covering grade-level math skills). We also believed that administering our own tests was important to address the problem of non-randomly missing tests in the CPS data, since lower-achieving students who have

higher rates of absence from school would be more likely to be absent on CPS testing day. In practice however this last potential advantage of the original in-person tests did not seem to be realized, in the sense that the pattern of missing-ness across the achievement distribution seems similar for the CPS and ISR tests.[17]

The achievement assessments we administered were those designed by the National Center for Education Statistics (NCES) for the 10th grade low-performing cohort in the first follow-up of the U.S. Department of Education's NELS:88 study. Of the 658 students who completed ISR surveys, 651 agreed to take the math test component of the survey, resulting in 641 valid test results (defined as answering 10 or more questions on the math assessment). We then had Educational Testing Service (ETS) run item response theory (IRT) models on these to allow us to create scale scores and also put our students onto the NELS:88 metric so that we can compare our scores to the NELS distribution nationwide.[18]

V. STUDY SAMPLE AND RANDOM ASSIGNMENT

In this section we describe how we selected the schools for our study and the students for our study within these schools, and then discuss how we carried out our random assignment.

A. Selecting Schools

Our research team worked with CPS to identify an initial list of schools that would be both large enough to enable us to randomize enough youth and fill up target program participation levels at each school, and would be likely to contain enough male students struggling academically who might benefit from programming. Our team identified 36 plausible high school partners. CPS then rank ordered these 36 schools based on their dropout rate, male arrest numbers, out of school suspension rates, and attendance rates to create a continuum of need for the intervention services, and recommended 30 schools to invite to a briefing with our research team in May 2013 during which we invited these schools to apply to participate.

[17] When we divide the study sample up into quartiles based on students' performance in the spring 2013 CPS test (the year before we randomly assigned youth and the intervention was delivered), we see a pattern of non-response across quartiles on this pre-test score that is similar for the CPS test data and the in-person ISR tests.

[18] ETS used a three-parameter IRT model (Lord, 1980) to put scores obtained on different sets of test items on the same scale for the purpose of comparisons in a common metric. ETS employed procedures both to ensure accuracy when converting the resulting raw item response data to final scores, and to ensure the accuracy and validity of the results, including: converting raw examinee item responses into scores for individual items; evaluating item functioning using both classical item analysis and IRT methods; and, assembling item data into meaningful and interpretable scores.

We then scheduled follow-up meetings with 27 schools to gauge their interest and capacity to participate in the study.[19] Of these, 6 declined to be considered for a variety of reasons, including scheduling and space issues, and concern about removing students from a regularly scheduled class or elective. Of the remaining 21 schools, 2 were being run by charter networks and did not have room for tutoring in their schedules, 2 more schools offered an International Baccalaureate program to their students and did not have room for tutoring in their schedules, 3 had student populations that were too small to support random assignment, and 2 were unable to commit before the required deadline, but later asked to be added into the study. We were ultimately able to offer Match in 12 high schools, primarily located on Chicago's south and west sides.[20]

B. Selecting Students

During the summer of 2013, we identified students within these 12 target CPS high schools who were eligible for the study using CPS administrative records from the 2012-13 academic year. Following the approach used in Heller et al. (2013), we first excluded those students who we thought were already too disengaged from school to attend regularly enough to benefit from a school-based program. This exclusion criterion was defined as having failed 75% or more of their classes in the previous school year *and* having missed more than 60% of their enrolled school days in that year. We also excluded students with serious disabilities as designated by the CPS data (autism, traumatic brain injury, emotional / behavioral disorder, educable mentally handicapped, and speech / language disabilities).

We then calculated an "academic risk index" that was a function of the number of prior-year course failures and unexcused absences, and being old for grade (interpreted as having been previously held back). Eligible students were then ranked on the basis of this risk index. We then determined the number of students we would need to randomly assign to fill up the program slots we had available in that school and chose that number of students in descending order on the ranked risk list. The share of male students selected to be eligible

[19] At the research briefing each school administration was asked to indicate their level of interest given the restraints of the study, and we then scheduled follow-up meetings with the 27 interested schools to answer questions about randomization, scheduling, space requirements, and the program providers. Principals were encouraged to invite any interested staff members to these meetings.
[20] Our schools (and the % of students free or reduced lunch eligible during the 2013-14 school year) are: Julian (91); Foreman (79); Harper (94); Harlan (83); Wells (94); Amundsen (89); Kelvyn Park (94); Marshall (87); MAS (96); World Language (93); Social Justice (92); Infinity (95).

in our study sample varies across schools because of school-by-school variation in both program capacity and school size. In practice, because of the scale of the experiment, in many schools we randomized all students who were not excluded based on their prior year course failures and absences. Essentially one can think of our study sample as a pool of male youth in distressed Chicago high schools in the middle of the distribution for these schools, with both the left (lowest achieving) and right (highest achieving) tails trimmed. The average test scores for our sample wind up being very similar to the school-wide averages in each study school (Table 1).

The one other complication in selecting our eligibility sample was that many students who were on the CPS rosters in summer 2013 as being expected to attend one of our study schools wound up not attending one of these study schools, either because they ended up choosing another school in CPS, their families moved or moved them to a private school, or they dropped out. In schools in which we wound up not having enough youth to randomly assign, we also added to the eligible sample new school entrants in the fall.[21]

We include every student we randomized and for whom we have CPS data in our study sample, including those we thought would be in our study schools when we randomized them in summer 2013 but then wound up not showing up at any of our study schools. In practice including or excluding those students does not affect our results much, since the probability that a youth randomized over the summer would show up in the fall at a study school seems to be balanced across the youth who were versus were not offered Match tutoring.

C. Random Assignment

All students in the study were independently randomly assigned to be offered the chance to participate or not in two programs, Match tutoring, to build academic skills, and the Becoming a Man (B.A.M) program, to address non-academic barriers to success. B.A.M. was developed and is implemented by Chicago nonprofit

[21] Schools began requesting randomization assignments as early as June 12, 2013, a full 10 weeks before the start of the 2013-14 school year. At that time, the 2012-13 school year had not finished, and so full academic year baseline data were not available to us when we began randomizing students. Instead, we used the same criteria (failing 75% of classes and missing more than 60% of enrolled days) on fall semester grades, and on attendance data through March 15, 2013. The other significant difference between the Heller et al. (2013) study randomization and ours is that the Heller et al. (2013) randomization took place after the 20th day of the school year when school enrollment is made "official" by the district, and presumably becomes more stable as fewer students transfer between schools after the 20th day. Our randomization, however, took place before the school year started, and was based on anticipated student rosters which we knew were only "best guesses" at who might actually enroll in the school. These rosters of 9th and 10th grade male students were provided to us by each of the study schools throughout the summer (beginning in June and extending into August). The timing of roster delivery to the study team was most often a function of each school's own internally determined readiness to begin setting student schedules for the upcoming school year.

Youth Guidance and aims to develop social-cognitive skills thought to be protective against violence involvement and anti-social behavior. Previous research has found that participation in B.A.M. reduces violent crime arrests and improves school engagement (Heller et al., 2013). The focus of the current study is to measure the effect of Match tutoring. We are able to isolate and identify this Match effect because of the way randomization was carried out. Just prior to the 2013-14 school year, eligible students were randomly assigned within each school to one of four experimental conditions: (a) control, (b) B.A.M. only, (c) Match tutoring only, or (d) B.A.M. and Match. As Figure 1 below shows, this creates a 2x2 factorial design. The focus of the present paper is to identify the main Match "row effect" (comparing average outcomes of cells C+D versus A+B).

Figure 1: Randomization Scheme - 2 x 2 Factorial Experiment Design

		B.A.M.	
		No	Yes
Match	No	A	B
	Yes	C	D

In order to accommodate the varying program capacity within each school, our random assignment algorithm varied the probability of treatment condition assignment.[22] Since our randomization was carried out separately by school and grade during the summer of 2013, we treat each school-grade combination as separate randomization blocks. In schools where too few students actually showed up in the fall for us to randomize, we identified new students entering the school (mostly during the first month of that school year) and randomly assigned them. For these students the randomization block is defined by the school and the time period in which the youth was randomized. All of our analyses below control for randomization-block fixed effects.

VI. ANALYSIS PLAN

In this paper, we present estimates of the effect of Match tutoring on student outcomes, with a primary focus in this draft on academic outcomes. Given the 2x2 factorial design of our experiment described in Figure 1 above, random assignment to B.A.M. and Match tutoring was independent. This means we can measure the effect of being offered Match tutoring by comparing the average outcome of students assigned to cells C+D,

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

who were offered Match tutoring, with the average outcomes of those assigned to cells A+B, who were not offered Match. The comparison of groups C to A captures the effect of being offered Match tutoring for students not offered B.A.M., while a comparison of group D to group B measures the effect of being offered Match for students who were offered B.A.M.

We present results from two types of analyses: estimates of the intent to treat (ITT) effect, and estimates of the effect of the treatment on the treated (TOT). The ITT estimate comes from estimating equation (1):

$$(1) \quad Y_{i1} = \pi_0 + \pi_1 Z_{i0} + X_{i0}\pi_3 + B_i\pi_4 + \varepsilon_{i1}$$

where Y_{i1} is an outcome for student *i* measured after random assignment, Z_{i0} is an indicator for having been randomly assigned to receive an offer to participate in Match tutoring (i.e., an indicator for having been randomly assigned to group C or D in Figure 1, as described above), B_i is a full set of randomization block fixed effects, ε_{i1} is a random error term, and X_{i0} is a set of baseline controls measured prior to random assignment that include test scores from the previous year and, in some models, also include age and grade fixed effects, free or reduced lunch status indicators, an indicator for having a learning disability, indicators for black and Hispanic, and the following academic measures measured in the 2012-13 school year: GPA, days absent, days of out-of-school suspension, days of in-school suspension, and the number of discipline-related incidents.

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 & 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.[23]

We demonstrate below that random assignment appears to have been carried out correctly, as evidenced by the fact that the distribution of baseline characteristics seems to be balanced across randomly assigned

[23] 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).

groups. The main other threat to internal validity in our study comes from selective attrition in our outcome measures, particularly in the CPS tests that students took in the spring of 2014. Fortunately, the likelihood that we have any CPS data for students during the intervention year is balanced across randomized groups, as is the likelihood of having CPS spring 2014 test scores. This means that selective attrition seems unlikely to be a major source of bias in our estimates, although as a sensitivity analysis we do also show results that use multiple imputation procedures to fill in missing outcome values. We also present estimates of a quantile regression on median test scores, imputing arbitrarily low scores (zeros) to students missing spring 2014 CPS tests given that the pre-randomization data suggest those with missing tests are drawn primarily from low achievers.

While the ITT estimates the effect of the offer of treatment, we might also be interested in the effect of participating in Match tutoring. To estimate the effect of participating in Match tutoring – the treatment on the treated (TOT) effect – we use 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 first-stage equation for the TOT estimation is:

(2) $D_{i1} = \gamma_0 + \gamma_1 Z_{i0} + X_{i0}\gamma_2 + B_{i0}\gamma_3 + \mu_{i1}$

where D is an indicator for having participated in Match tutoring, which we define as having participated in at least one Match tutoring session, the $\gamma's$ are parameters to be estimated, μ is a random error term, and all other variables are defined as above. The relationship of interest is:

(3) $Y_{i1} = \beta_0 + \beta_1 D_{i0} + X_{i0}\beta_3 + B_i\beta_4 + \upsilon_{i1}$

where υ is a random error term and the $\beta's$ are parameters to be estimated. 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 Table 2).

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 41 percent for all youth assigned into either treatment arm, the TOT estimate will be about 1/.41 = 2.44 times the ITT.

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.

In this draft of the paper we present p-values all calculated from pairwise comparisons, that is, ignoring the number of different impact estimates we calculate other than our decision to limit the number of outcomes we examine and divide outcomes up into pre-specified families or domains. In the next draft of this manuscript we will also present the results of accounting for multiple testing issues following the approach from Cook et al. (2014) – see also Kling, Liebman and Katz (2007) and Anderson (2008).

VII. MAIN RESULTS

In this section we present our main impact estimates for the effects of the intensive individualized instruction delivered by Match on student schooling and behavioral outcomes. The estimated effects of treatment on the treated (TOT) for math achievement are on the order of 0.19 to 0.31 standard deviations, depending on the exact test and norming that we use. These impacts are measured by the ACT Inc.'s EXPLORE and PLAN tests, which CPS administers to 9th and 10th graders, respectively, as well as in-person achievement tests that our survey subcontractor (the Institute for Social Research at the University of Michigan) administered

to a randomly selected sub-sample of our larger study sample. The fact that we see similarly sized impacts on the math achievement test administered by our survey subcontractor, a test that is different from the focal high-stakes test administered by CPS, is one indication that the results we see on the EXPLORE and PLAN tests are not the result of a narrow "teaching to the test" by the Match tutors. A similar conclusion is suggested by the fact that we see very large improvements in math grades (0.45 to 0.49 SD) and sizable reductions in course failures in math (between one-half and two-thirds of the control mean) and course failures overall (about one-quarter of the control mean).

A. Descriptive Statistics

Table 1 provides some initial context for our study sample. We show the average percentile rankings for 9th and 10th graders pooled together on the EXPLORE/PLAN math tests during the post-random assignment period (spring 2014). We show these average scores for the entire CPS school system (first row), all students attending our study schools (second row), our study sample specifically within these study schools – that is the set of students we randomly assigned (we have valid post-randomization test scores for 1,919 of our study sample of 2,718, as shown in the third row), and finally the set of students we randomized and for whom we have post-randomization (spring 2014) test scores and were tested in our 12 study schools (1,646).[24]

The first striking observation in Table 1 is that the average percentile ranking for the CPS system is the 50th percentile. In interpreting this result it is important to keep in mind that CPS administers the 9th grade EXPLORE and 10th grade PLAN tests in the spring, so that CPS can measure year-to-year growth in test scores through the 11th grade administration of the ACT test, while the percentile scores for the 9th grade EXPLORE tests, which are provided to CPS by ACT, are normed to the national distribution of scale scores from students who took the EXPLORE in the fall, not the spring. This means CPS 9th graders have almost a full additional academic year of schooling more than the norming population to whom they're being compared.[25] This issue should not affect our estimates for the effect of treatment-control differences from Match tutoring because both treatment and control students are normed in the same way. Furthermore, though the level of the percentile

[24] Students that we randomly assigned but then transferred or moved to another CPS school are still included in our estimation sample – exclusion of that group of students is the difference between rows 3 and 4 in Table 1.
[25] See footnote 17.

score may be hard to interpret, the difference in the percentile scores between the treatment and control group is probably a reasonable approximation of the size of the treatment effects in percentile points. We also report results for normalized scaled scores that do not rely at all on the percentile norming.

Table 1 shows that baseline math test scores in the 12 study schools we examine here were about 12 percentile points lower on average than the CPS system-wide average. This is consistent with our intentional selection of relatively more disadvantaged schools within the CPS system. Most of the study schools are located on the economically and racially segregated south and west sides of Chicago. It is worth noting the baseline test scores varied fairly significantly across the schools in the study, ranging from an average of 24.7th percentile at Harper High School (the site of our earlier pilot study) to 58.8th percentile at Infinity High School. Within each school, the average baseline test scores of our study sample are close to the school-wide averages. This is not surprising since, as described above, our sample selection process consisted of trimming the extreme left and right tails of the achievement distribution and randomly assigning male students in the rest of the distribution in these disadvantaged high schools.

Table 2 presents additional baseline statistics for our study sample. Almost all youth in our sample are either African-American or Hispanic, about evenly split between the two groups. Over 90% are eligible for free or reduced price lunch. The average GPA for these youth during the year before our intervention (that is, during the 2012-13 school year) was about 2.1 on a 4-point scale. On average youth in our study sample missed about a month of school during the pre-program year and were suspended from school for nearly 2 days.

Table 2 also shows that random assignment appears to have been carried out correctly, in the sense that the distributions of baseline characteristics seem to be balanced between youth randomly assigned to receive Match (that is, cells C and D in the 2x2 table above – youth assigned to either Match only or B.A.M. + Match) or not (cells A and B in the 2x2 table above, youth assigned to receive either B.A.M. only or no services through the study). Only one of the pairwise differences in baseline characteristics is significant even at the 10% level; these pairwise tests come from regressions of the baseline variable on an indicator for having been assigned the offer of Match, controlling for randomization block fixed effects. We also carry out an omnibus F-

test of the null hypothesis that the baseline characteristics are jointly the same across the Match treatment and control groups by regressing a treatment-group indicator against all the variables in Table 2 (also controlling for randomization blocks given our design described above). The p-value on that joint test is p=.26.

The tables show additional results for different baseline balance tests carried out separately just for the youth we randomly assigned during the summer of 2013 (those youth who we thought would be attending our study schools in the fall) and with the new entrants to our schools who we randomly assigned during the fall of 2013. The p-values for these F-tests equal 0.19 (N=2,219) and 0.67 (N=499), respectively.

B. Match Impacts

Before we present the results for Match tutoring on student math achievement and other outcomes, it is worth noting again that the sample for whom we have valid post-randomization (spring 2014) CPS test scores on the EXPLORE and PLAN assessments is doing substantially better in school than those for whom we are missing spring 2014 test scores, as shown in Table 3. Luckily the rate of CPS test missing-ness is similar for the group of youth assigned to be offered Match or assigned to our control group (not offered Match). Similarly the differences in average baseline characteristics between those for whom we do versus do not have spring 2014 CPS tests are also similar between our Match treatment and control groups.

Table 4 shows our main results for the Match effect (the "row comparison" in our 2x2 table, that is the average outcomes for cells C and D versus the average outcomes for cells A and B). We show the control mean in the tables as well as both the ITT and TOT effects. Some readers will prefer the ITT results because they require no assumptions other than that random assignment was carried out correctly. Other readers will prefer the TOT estimates because they can be useful in drawing inferences about what effects might be in other contexts where take-up rates are different (so long as there is not an excessive amount of treatment heterogeneity). The TOT also helps in carrying out benefit-cost analyses of the sort most people are used to, that is comparing benefits per participant with costs per participant. In Table 6 we also show what role the baseline covariates play in our analysis, by showing results that control only for randomization-block fixed effects, as

well as models that add in various combinations of baseline covariates, including a full set of baseline covariates as described above. All standard errors are clustered at the school level.

Because standardized achievement tests are perhaps the easiest outcome to compare across studies (since different school districts may vary with respect to grading standards or attendance rules), we begin our discussion with those results. We focus on the TOT result to simplify the exposition, but readers who prefer the ITT result instead should feel free to focus on that column in the tables.

Table 4 shows that the TOT test score gain in math scores in the CPS data ranges from 0.19 to 0.23 SD if we use the scale score version of the EXPLORE/PLAN, and normalize to the national distribution or the control group distribution, respectively, to convert these into Z-scores. The TOT effect in percentile terms is nearly 8 percentile points. While the percentile rank version of the scores is a monotonic transformation of the scale score version of the test results, the percentile rank format "shrinks" and "stretches" different parts of the distribution differently, which explains why the t-statistics are slightly different across the two types of test scalings. The Z-score effect with the percentile rankings ranges from 0.27 to 0.31 SD, depending on whether we normalize using the national distribution for these tests or instead by the control group's distribution. These results are highly statistically significant whether we use standard errors clustered at the school level or instead use a permutation test to calculate p-values.

A potential concern with these results is that they could simply reflect that the Match tutors are "teaching to the test," rather than improving the amount of broad-based math knowledge or skills of youth. It is for that reason somewhat reassuring that we also find large improvements in math GPA – that is, the CPS math teachers themselves also see sizable gains in math performance among our study sample. Table 4 shows TOT effects on math grades equal to 0.58 points on a 1-4 grade scale, a sizable gain compared to the control complier mean math grade point average of 1.77. In Z-score terms this equals about 0.50 SD. We also see in Table 4 that Match tutoring reduces the number of math course failures per youth by about 0.19, which equals about one half percent of the control complier mean.

The next panel shows Match impacts on subjects outside of math. As a reminder, the Match tutoring focused only on math, so this panel examines the degree to which there are spillovers from improved math achievement into other outcome domains. While we do not see any gains in scores on the EXPLORE/PLAN reading assessments, or on overall GPA in non-math classes, we do see a reduction in total course failures.[26]

The next two panels examine Match effects on different behavioral outcomes, first measured by CPS data and then measured by CPD arrests (although as a reminder the current set of CPD arrests only run through July 1, 2014, and so may miss much of the impact on criminal behavior, which tends to be so disproportionately concentrated over the summer). While we see no statistically significant effects on any behavioral measure in the CPS data, we do see proportionately large negative point estimates for the number of arrests youth experienced for violent crimes (over 50% of the control complier mean, $p<.10$). Estimated treatment effects on property, drug, and other arrests are insignificantly different from zero. We are working to update these arrest data to cover the rest of the summer, which we think will provide improved statistical power and a more accurate assessment of the intervention's effect on youth behavioral outcomes.

A different type of concern with the CPS test results is that we are missing spring 2014 CPS test data for 799 of the 2,718 youth in our main analysis sample (29%). Table 5 reports the results of our analysis of the original math achievement test results that ISR administered to a randomly selected sub-sample of youth in our experiment. (We tried to survey and test only a sub-sample rather than the full sample because of budget constraints). The tests that we administered were designed by the NCES for the U.S. Department of Education's NELS:88 nationally representative study.

The TOT effect on these tests (standardized to the control group distribution) is equal to 0.295 SD. This is towards the upper end of the range of Z-score impacts that we see from the tests administered by CPS, which runs from 0.19 to 0.31 SD. The fact that we see such changes on a test different from the one that CPS administers also provides some evidence against the interpretation of our findings as merely the result of teaching to the CPS test.

[26] Because the 10th graders in our study take Match tutoring instead of an elective, to avoid some mechanical effect of Match on the number of non-math course failures we focus just on grades in non-math core classes. The average number of these classes taken by treatment and control youth is balanced.

Table 6 presents sensitivity analyses for our main results, showing how our ITT impacts in the CPS administrative data change as we vary the set of baseline controls that we include in the estimation model. The table shows that most of our findings are qualitatively similar regardless of whether or not we condition on baseline attributes of our youth, or what specific baseline characteristics we control for. The next draft of the paper will also show how the results change when we use multiple imputation to fill in missing outcome values, and of course we will also present p-values that adjust for multiple comparisons.

Table 7 shows the Match impacts on different survey questions that ISR asked to a random subset of our study sample, to try to get at candidate mechanisms. In principle one potential pathway through which the intervention might operate is as a mentoring program. That is, the intervention might change outcomes by increasing the degree to which youth feel connected to a pro-social adult – what Coleman (1988) called "social capital" – since the small-group tutoring in Match might have made it easier for the tutor to build rapport with the students, in comparison to the challenge regular teachers face when trying to do this in a full classroom with many more students. Yet we see no statistically significant changes in the degree to which youth report feeling like they have adults who they are comfortable talking to about personal problems, or who care about how their lives will turn out, or who they can talk to if they need help with school work. Nor do we see any evidence that the program changed the degree to which youth are spending time with more pro-social peers or peers who take school more seriously. If social capital is actually an important mechanism, it must be operating through some more subtle pathway that we have not been able to measure with our surveys.

The survey data that we have suggests that the academic focus of the intervention itself seems to be an important part of the mechanism of action. We find some support for the idea that the individualized instruction increases self-confidence and the degree to which students are inclined or able to be engaged with school. Interestingly, this channel seems to be mostly narrowly targeted to youth attitudes towards math. Youth who participate in the program are no more likely than controls to say that they like school, spend more time on homework in general, or have higher expectations or aspirations for their schooling attainment. The main change is that participants are more likely to say that they like math (0.25 SD is the TOT effect, $p<.10$) and get

good grades in math (0.38 SD, $p<.05$). The only other indicator of youth attitudes towards school that changes in our data is the degree to which youth report that good grades are important to them (TOT of 0.23 SD, $p<.10$).

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 instead focus more resources on academic interventions in early childhood (for example Carneiro & Heckman, 2003).

The impacts from the large-scale randomized experimental intervention that we report on here 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 deficits in academic skills that can for some youth be very sizable and leave them needing help with material that is many years below grade level. This in turn leads to a mismatch between what many students who are falling behind need and what regular school settings deliver. Have previous interventions for disadvantaged youth mostly been aiming at the wrong target?

We find that an intervention that delivers a high dose of individualized instruction – daily math tutoring delivered within the school day in a way that holds down costs – seems to generate large gains in learning with a sample of low-income male youth attending CPS high schools in economically disadvantaged neighborhoods in the city of Chicago. The observed gain on a broad-based test in math achievement administered by the CPS system (the ACT Inc's EXPLORE and PLAN tests) equals 0.19 to 0.31 of a SD, depending on what test score scaling we use and what population we use to norm the test. One indication that this is not simply the result of narrowly teaching to the test is that we also observe large reductions in math course failures and increases in

math GPA. We also had our survey subcontractor administer to youth an entirely different test (the assessments developed for the NELS:88 study), where we observe gains equal to about 0.30 SD.

As one way to judge the magnitude of our test score result, the effect in percentiles measured relative to the nationwide test score distribution (0.27 SD) is equal to one-third 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 this much, since the effects could be different for different populations, and in particular, we have no idea at present how cohorts of predominantly 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 ours is not a very expensive intervention and that the duration of exposure to the intervention was only one academic year.

A general concern in the field of education is whether interventions can succeed at large scale. The current paper reports results from a study that randomized about 2,700 male students across 12 CPS high schools. This is equal to about 5% of all male high school students in the city of Chicago. As indicated in Tables 1 and 2, these youth are performing below average within the CPS system in school, so our sample is not "cherry picked" from a pool of easy-to-work-with students. Most youth in the study come from very economically disadvantaged and racially segregated neighborhoods, although the study did include a few high schools that are located in less distressed north-side neighborhoods of the city. The sample also includes large numbers of both African-American and Hispanic youth. We worked closely with CPS to identify a candidate list of 36 disadvantaged schools, and then asked CPS to rank order the schools by need based on dropout rate, male arrest rates, out-of-school suspension rate, and attendance, prioritizing those schools ranked highest.

There is also a question about what the scarce inputs are in delivering Match tutoring that might prevent the intervention from being delivered at even larger scale in the future. While this is an important question for future research, it is worth noting that the intervention has been delivered at large scale in a number of cities (although this is the first randomized experimental test of the intervention). Match tutoring has been delivered to over 700 students per year across three schools in Boston; 500-600 students in several public high schools in

Lawrence, Massachusetts; and 3,000 students across 13 schools in Houston (Fryer, 2014). 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) 8 applications per opening in Chicago.

The impacts we report here are quite sizable (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. For example, the impact in math test scores that we observe for disadvantaged 9th and 10th graders who receive just Match tutoring is at least as large as the effect Fryer (2014) reports from providing younger children (4th graders) with a combination of this type of tutoring *plus* a mix of other "no-excuses" charter school principles that were injected into the schools (from 0.19 to 0.31 SD in our study, versus 0.22 SD in Fryer 2014, Table V).

Our best estimate for the cost per participant in our intervention is roughly $3,800, with a defensible range of $3,500 to $4,300.[27] The test score gain per dollar spent from this intervention is very large compared to previous interventions for disadvantaged youth; aside from Fryer's (2014) 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.[28]

Figure 2 compares our test-score impacts per dollar spent to several other commonly cited studies in the literature that have been carried out with younger children. The pathway through which early-childhood interventions affect long-term outcomes is not well understood, and does not seem to be mediated always by persistent impacts on adolescent test scores. (That is, we see some interventions that change test scores in the short term during early childhood, then those test-score gains fade away, yet we see long-term impacts on adult

[27] Our realized cost was about $4,300 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 $3,500 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 Match tutoring intervention studied at scale by Fryer (2014) is reported to be $2,500 per student.

[28] While a few interventions have been shown to boost high school graduation rates for youth (see Krueger, 2003a; Guryan, 2004; Bloom, Muller-Ravett & 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.

outcomes.) So we focus on comparing impacts measured at the end of the intervention period across interventions, regardless of how old children were at the time the impacts were measured.

Figure 2 shows that the estimated impacts (on math test scores measured at the end of the intervention period) per dollar spent for the high-dosage tutoring intervention we studied in Chicago seem to be at least as large as those from other widely cited educational interventions that have been rigorously studied, including those that are delivered in early childhood. For example while the "raw" overall effect of Perry Preschool is much larger than what we see with our high-dosage tutoring intervention (Schweinhart et al., 2005),[29] Perry is also substantially more expensive and so winds up having somewhat smaller impacts per dollar spent compared to the intervention for disadvantaged teens that we study here. The same general pattern holds if we compare impacts per dollar spent to the effects on test scores from cash transfers from the Earned Income Tax Credit, or EITC (Dahl & Lochner, 2012), class-size reduction in early elementary school grades (Krueger, 1999, 2003b; Schanzenbach, 2006), or Head Start (Phillips & Ludwig, 2007, 2008).

Our findings 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 grade-level material. Efforts to address this mismatch 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 is the recognition that the tutoring method of instruction substantially changes the set of skills and experiences required to succeed as an instructor. There is, in short, the possibility of a tradeoff between the "teaching technology" used to deliver academic instruction and teacher "quality," as 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 is

[29] The only short-term test scores available in Perry Preschool are IQ scores, so we present in Figure 2 impacts in effect size terms per dollar spent at the end of the intervention period.

that the need for helping students who fall behind might be only temporary. Our results suggest that youth who receive the Match tutoring learn about an *extra* 1 to 2 years' worth of math above and beyond what the control group learns over the course of the academic year.[30] In contrast, students who are 4 to 10 years behind grade level, as unfortunately is not uncommon in distressed urban areas, are in a position where they have been getting very little or virtually nothing out of regular classroom instruction for years. If it is possible to achieve at large scale the results we report here, 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 with and benefit from the grade-level material taught in regular high school classrooms.

Perhaps the most important and broad lesson to come from these findings is with respect to the value of intervening at different points in the life course, and in particular, the possibility of substantially and cost-effectively changing the academic outcomes of disadvantaged children even once they reach adolescence so long as the right (individualized) intervention strategy is used. 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 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, especially now that we may be diagnosing the key underlying problems.

[30] Reardon (2011, p. 97) reports that in the NAEP, the average U.S. student gains between 0.60 to 0.70 standard deviations on the NAEP test scores between 8th and 12th grade, or about 0.15 to 0.175 SD per year.

REFERENCES

Anderson, M. L. (2008). 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, 1481-1495.

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

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007) Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics,* 122(3), 1235-1264.

Barrow, L., Claessens, A. & Schanzenbach, D. W. (2013). *The impact of Chicago's small high school initiative*. Northwestern University, Institute for Policy Research Working Paper, WP-13-20.

Bloom, B. 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, D., Muller-Ravett, S., & Broadus, J. (2011). *Staying on Course: Three-year results of the National Guard Youth ChalleNGe Evaluation*. New York, NY: MDRC.

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

Carneiro, P. & Heckman, J. (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, E. U., & Staiger, D. O. (2012). *Knowledge, tests, and fadeout in educational interventions*. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 18038.

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

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94, S95-S120.

Cook, P. J., Dodge, K., Farkas, G., Fryer Jr, R. G., Guryan, J., Ludwig, J., Mayer, S., Pollack, H. & Steinberg, L. (2014). *The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth: Results from a Randomized Experiment in Chicago.* Cambridge, MA: National Bureau of Economic Research, Working Paper No. 19862.

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

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

Duflo, E., Dupas, P. & Kremer, M. (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, G., Dowsett, C.J., Claessens, A., Magnuson, K., Huston, A.C., Klebanov, P., Pagani, L., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428-1446.

Duncan, G. J. &Murnane, R. J., Eds. (2011). *Whither Opportunity? Rising Inequality, Schools, and Children'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, B. & Tibshirani, R. J. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57.

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.

Fryer, R. G. (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. *Quarterly Journal of Economics*, 129(3), 1355-1407.

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

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

Guryan, J. (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.

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

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

Jencks, C. & Phillips, M. Eds. (1998). *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

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

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1), 83-119.

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

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

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Ludwig, J. & Phillips, D. A. (2007). *The benefits and costs of Head Start*. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 12973.

Ludwig, J. & Phillips, D. A. (2008). The Long-Term Effects of Head Start on Low-Income Children. *Annals of the New York Academy of Sciences*, 40, 1-12.

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

Nomi, T. & Allensworth, E. (2009). Double-dose Algebra as an alternative strategy to remediation: Effects on students’ outcomes. *Journal of Research on Educational Effectiveness*, 2(2), 111-48.

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

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

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R. & Nores, M. (2005). *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40*. Ypsilanti, Michigan: High/Scope Press.

Figure 2:
Comparison of test score impacts per $1,000 of per child spending

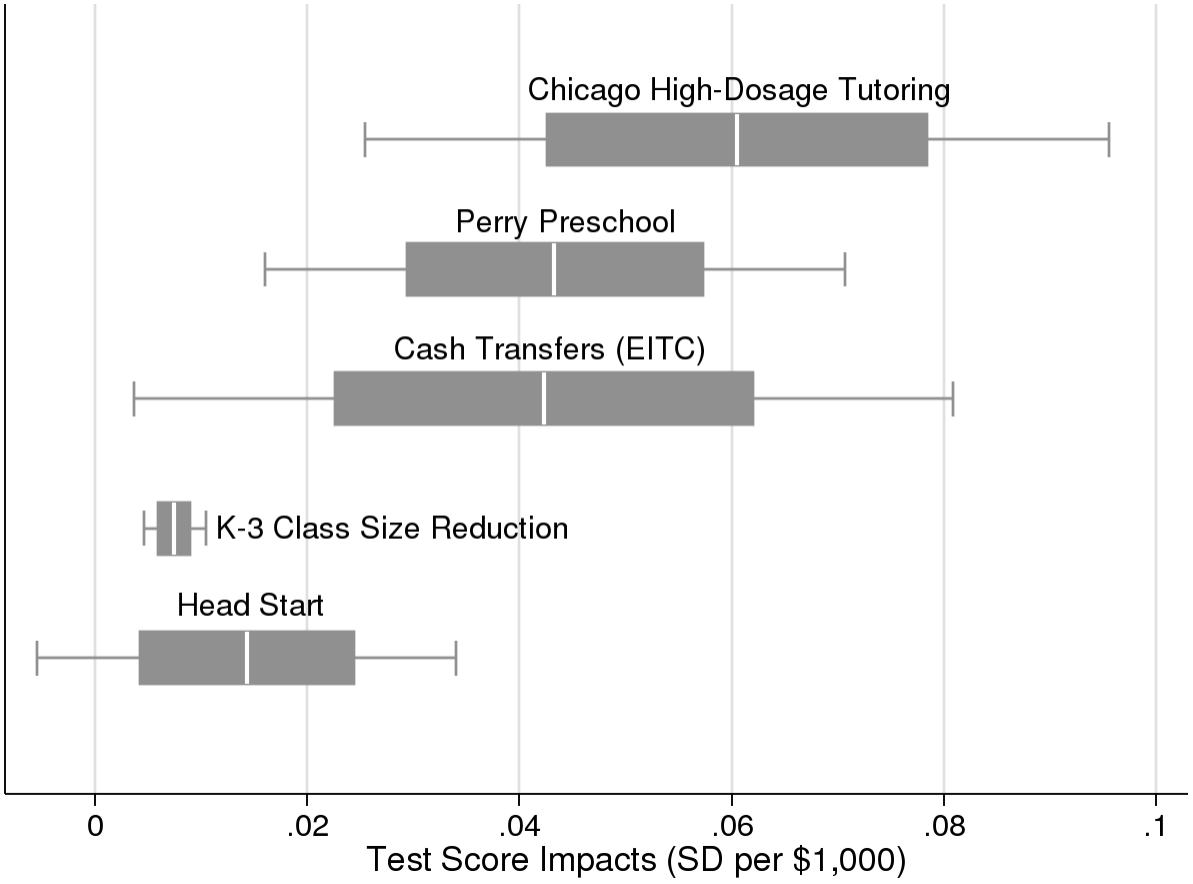

Table 1. SY2013-14 Explore/Plan National Percentile Ranking

	Average score	25th percentile	Median	75th percentile
All CPS Students (n = 48,193)	49.71	23	43	77
All students in study schools (n=3,510)	37.318	19	33	54
Students randomized into study schools (n=1,919)	39.513	19	33	61
Students randomized and tested in study schools (n=1,646)	39.938	19	33	61

Students randomized into study schools : all students in randomized into our BAM+Match study without regard for which school they took the test in. If a student was randomized into school A (study school) but took the test in school B (not a study school) their test scores from school B are included here.

Students randomized and tested in study schools: students randomized into BAM+Match schools who show up in the EPAS data as having taken the test in a BAM+Match school.

Table 2. Baseline means and balance test for all students randomized into study (N = 2,718)

				treatment-control difference (p-value)
	All (N = 2,718)	Mean for Match = 0 (N = 1392)	Mean for Match = 1 (N = 1326)	Covariates: p > F = 0.2409
N Students	2718	1392	1326	
Age 14 on Sep 1, 2013	35.95%	35.78%	36.12%	0.807
Age 15 on Sep 1, 2013	45.44%	46.34%	44.49%	0.613
Age 16 on Sep 1, 2013	15.34%	15.09%	15.61%	0.419
9th grade during SY2013-14	56.11%	55.75%	56.49%	0.222
10th grade during SY2013-14	43.23%	43.46%	42.99%	0.221
"Free lunch" noted on student's CPS record	86.87%	87.43%	86.27%	0.971
"Reduced lunch" noted on student's CPS record	4.38%	4.96%	3.77%	0.215
Black	47.09%	46.26%	47.96%	0.724
Hispanic	47.76%	48.78%	46.68%	0.659
Other race	5.15%	4.96%	5.35%	0.373
"Learning disabilty" noted on student's CPS record	17.22%	17.03%	17.42%	0.341
School Year 2012-13 Total GPA	2.10	2.12	2.08	0.579
School Year 2012-13 Total courses failed	2.32	2.32	2.31	0.076 +
Number of absences, SY2012-13	21.24	20.77	21.74	0.828
Number of discipline reported incidents, SY2012-13	1.30	1.29	1.31	0.903
Days out-of-school suspension (from incident data), SY2012-13	1.69	1.68	1.70	0.897
Spring 2012-13 Explore/Plan math score, national percentile	39.13	39.62	38.61	0.167
Spring 2012-13 Explore/Plan reading score, national percentile	37.73	38.11	37.33	0.463
Arrest history before August 26, 2013	18.25%	18.39%	18.10%	0.358

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

Mean for control displays means for students assigned to the control group. **Match= 0** contains all students assigned to BAM only or contol. **Match = 1** contains all students assigned to BAM+Match or Match only. **P-value** for mean comparison from ANOVA.

Treatment-control difference column shows the p-values for different ways of testing the significance of treatment-control differences. **Full model** test gives the model f-test results from an OLS regression where BAM+Match/Match Only group assignment is the outcome and all the baseline measures listed above are the right hand side variables. The model also controls for randomization block fixed effects. The test of **covariates** gives the results from a post-estimation f-test of whether all the baseline covariates are jointly equal to 0. **Individual row p-values** are the result of an OLS model where the baseline is the outcome, treatment assignment is the right hand variable. Models control for randomization fixed effects with standard errors clustered in schools.

Table 3: Match Baseline means by randomized group and availability of spring 2014 post-tests in math

	dmatch=0 has spring test	dmatch=0 no spring test	dmatch=1 has spring test	damtch=1 no spring test
N Students	N = 986	N = 406	N = 933	N = 393
Age 14 on Sep 1, 2013	0.39	0.29	0.39	0.30
Age 15 on Sep 1, 2013	0.48	0.41	0.48	0.36
Age 16 on Sep 1, 2013	0.12	0.23	0.12	0.25
9th grade during SY2013-14	0.54	0.60	0.55	0.60
10th grade during SY2013-14	0.46	0.37	0.45	0.38
"Free lunch" noted on student's CPS record	0.89	0.84	0.88	0.83
"Reduced lunch" noted on student's CPS record	0.05	0.04	0.04	0.03
"Learning disabilty" noted on student's CPS record	0.17	0.17	0.18	0.16
School Year 2012-13 Total GPA	2.32	1.60	2.28	1.61
School Year 2012-13 Total courses failed	1.61	4.14	1.53	4.19
Number of absences, SY2012-13	15.12	34.78	15.96	35.57
Number of discipline reported incidents, SY2012-13	0.77	2.58	0.91	2.25
Days out-of-school suspension (from incident data), SY2012-13	0.90	3.62	1.09	3.14
Arrest history before August 26, 2013	0.11	0.37	0.11	0.36
Missing Spring Sy2013-13 Test math score	0.09	0.37	0.13	0.32
Spring 2013-14 Explore/Plan math score, national percentile	37.97		41.14	
Spring 2013-14 Explore/Plan reading score, national percentile	35.61		35.71	

Students participated in the intervention during AY13-14. Student outcomes for AY12-13 occur the year before the program.

Table 4: Estimated effects of program offer and participation on student learning outcome and behavior during program year (SY13-14)

	Control mean (dmatch = 0)	Intent to Treat (ITT)	Treatment on Treated (TOT)	Control Complier Mean	permutation test p-value
Outcome Domain: Math Achievement					
Math Achievement Test Scores Spring 2014 (Explore/Plan), N = 1,919					
Scale score	14.867	0.439**	0.807***	14.397	0.0003
		[0.141]	[0.223]		
Scale Score (Control Distribution)	-0.064	0.125*	0.230***	-0.213	0.0004
		[0.043]	[0.068]		
Scale Score (National Distribution)	-0.525	0.102*	0.188**	-0.640	0.0009
		[0.037]	[0.057]		
National Percentile Rank	37.974	4.262**	7.843***	33.738	0
		[1.114]	[1.725]		
National Percentile Rank (Control Distribution)	-0.045	0.169**	0.311***	-0.213	0
		[0.044]	[0.068]		
National Percentile Rank (National Distribution)	-0.417	0.148**	0.272***	-0.564	0
		[0.039]	[0.060]		
Math GPA 2013-2014, N = 2,286					
Math GPA 2013-2014 (1-4 point scale)	1.771	0.295***	0.581***	1.588	0
		[0.047]	[0.093]		
Math GPRA Z Score (Control Distribution)	-0.029	0.253***	0.498***	-0.185	0
		[0.040]	[0.080]		
Math Courses Failed 2013-2014, N = 2,286					
Math Courses Failed 2013-2014	0.371	-0.098**	-0.192**	0.387	0.0003
		[0.030]	[0.059]		
Outcome Domain: Reading Achievement in Other (Non-Math) Subjects					
Reading Achievement Test Scores Spring 2014 (Explore/Plan), N = 1,918					
Scale score	13.533	0.058	0.106	13.203	0.6714
		[0.100]	[0.173]		
Scale Score (Control Distribution)	-0.038	0.022	0.041	-0.180	0.5624
		[0.034]	[0.058]		
Scale Score (National Distribution)	-0.577	0.014	0.026	-0.665	0.5808
		[0.023]	[0.039]		
National Percentile Rank	35.623	0.384	0.707	32.773	0.6385
		[0.665]	[1.147]		
National Percentile Rank (Control Distribution)	-0.039	0.017	0.032	-0.168	0.6404
		[0.030]	[0.052]		
National Percentile Rank (National Distribution)	-0.498	0.013	0.024	-0.551	0.6409
		[0.023]	[0.040]		
GPA in Non-Math Courses 2013-2014, N = 2,312					
Non-math GPA, SY2013-14	1.859	0.072*	0.143**	1.722	0.0208
		[0.026]	[0.050]		
Non-math GPA, z-score in control group SD's	-0.017	0.069*	0.137**	-0.148	0.0207
		[0.025]	[0.048]		
Non-Math Courses Failed 2013-2014, N = 2,312					
All non-math course failures, SY2013-14	2.108	-0.347**	-0.687***	2.457	0.0004
		[0.102]	[0.199]		
All core non-math course failures, SY2013-14	1.288	-0.165*	-0.328*	1.519	0.0203
		[0.065]	[0.127]		
Outcome Domain: Behavior					
Discipline Incidents SY13-14, N = 2,575	1.505	0.066	0.146	1.551	0.5281
		[0.145]	[0.299]		
Days Absent SY13-14, N=2,575	24.173	0.543	1.196	23.204	0.5001
		[0.922]	[1.885]		
Out-of-School Suspension Days, N = 2,575	1.507	0.178	0.392	1.593	0.2303
		[0.216]	[0.435]		
Outcome Domain: Crime					
Arrests for Violent Crimes Aug. 26, 2013--Jul. 1, 2014, N = 2,718	0.039	-0.016	-0.038+	0.062	0.0383
		[0.010]	[0.021]		
Arrests for Property Crimes Aug. 26, 2013--Jul. 1, 2014, N = 2,718	0.062	-0.004	-0.009	0.046	0.7433
		[0.011]	[0.023]		
Arrests for Drug Crimes Aug. 26, 2013--Jul. 1, 2014, N = 2,718	0.05	0.013	0.031	0.006	0.2693
		[0.014]	[0.032]		
Arrests for Other Crimes Aug. 26, 2013--Jul. 1, 2014, N = 2,718	0.19	-0.03	-0.071	0.189	0.1681
		[0.019]	[0.044]		

+ p < 0.1; * p < 0.05; ** p<.01; *** p<.001

Model information: randomization block number fixed effects included in ALL models. Standard errors clustered within schools. All students with non-missing outcome data for each measure are included in the models

Baseline Demographics: indicator for age 14, indicator for age 15, indicator for 10th grade, indicator for free lunch recipient, indicator for learning disability, indicator for black, indicator for hispanic,

Baseline 2012-13 academic measures: Overall GPA , number of absences, and Explore/Plan reading and math scaled score (standardized to control group). For each of these variables, missing observations were recoded as 0's and indicators such as "missing 2012-13 GPA" were created and included in the models.

Baseline 2012-13 school behavioral measures: number of days in out-of-school suspension, number of discipline incidents. Like the academic measures, missing observations were recoded as 0's and indicators were created and included in the models.

Crime outcomes include the following additional baseline measures: number of prior violent arrests , number of prior property arrests, number of prior drug arrests, number of prior arrests which are not violent, drug, or property related.

Permutation p-values 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

Table 5. Match ITT and TOT effect for CPS and ISR tests

	Control mean (dmatch = 0)	ITT		TOT	
		b	*se*	*b*	*se*
Explore/Plan Math Results (N=1,919)					
Math score, scaled	14.867	**0.439****	[0.141]	**0.807*****	[0.223]
Math score, scaled, z-score (control distribution)	-0.064	**0.125***	[0.043]	**0.230*****	[0.068]
Math score, scaled, z-score national distribution)	-0.525	**0.102***	[0.037]	**0.188****	[0.057]
Math score, national percentile	37.974	**4.262****	[1.114]	**7.843*****	[1.725]
Math score, national percentile, z-score (control distribution)	-0.045	**0.169****	[0.044]	**0.311*****	[0.068]
Math score, national percentile, z-score (national distribution)	-0.417	**0.148****	[0.039]	**0.272*****	[0.060]
ISR achievement tests(N = 641)					
NELS 88 Scale Score	42.13	**2.10****	[0.730]	**3.36****	[1.167]
NELS 88 Scale Score, z-score (control distribution)	0	**0.182****	[0.063]	**0.291****	[0.101]

p < 0.1 +
p < 0.05 *
p < 0.01 **
p < 0.001 ***

Academic outcome models use randomization block number fixed effects and demographic, schooling and behavioral covariates. Standard errors clustered within schools.

Mean for ISR outcomes are weighted by ISR survey weights. All ISR models use desgin-based standared errors calculated using Taylor Series linearization.

Table 6: Sensitivity Analyses, controlling for different combinations of baseline covariates

	ITT - Full covariates	ITT controlling only for socio-demographics	ITT controlling only for prior schooling outcomes	ITT controlling only for prior criminal factors	ITT - No covariates
Outcome Domain: Math Achievement					
Math Achievement Test Scores Spring 2014 (Explore/Plan), N = 1,919					
Z Score (Control Distribution)	0.125*	0.101+	0.119*	0.086	0.086
	[0.043]	[0.051]	[0.047]	[0.055]	[0.054]
Z Score (National Distribution)	0.102*	0.083+	0.099*	0.072	0.072
	[0.037]	[0.044]	[0.039]	[0.046]	[0.046]
National Percentile Rank	4.262**	3.705*	4.080**	3.328*	3.339*
	[1.114]	[1.207]	[1.190]	[1.304]	[1.276]
Math GPA 2013-2014, N = 2,286					
Math GPA 2013-2014 (1-4 point scale)	0.295***	0.279***	0.299***	0.292***	0.285***
	[0.047]	[0.050]	[0.047]	[0.057]	[0.055]
Math GPA Z Score (Control Distribution)	0.253***	0.239***	0.256***	0.250***	0.244***
	[0.040]	[0.043]	[0.040]	[0.049]	[0.047]
Math Courses Failed 2013-2014, N = 2,286					
Math Courses Failed 2013-2014	-0.098**	-0.091*	-0.101**	-0.101*	-0.099*
	[0.030]	[0.033]	[0.032]	[0.038]	[0.038]
Outcome Domain: Reading Achievement in Other (Non-Math) Subjects					
Reading Achievement Test Scores Spring 2014 (Explore/Plan), N = 1,918					
Z Score (Control Distribution)	0.022	0.004	0.012	-0.017	-0.014
	[0.034]	[0.047]	[0.036]	[0.051]	[0.049]
Z Score (National Distribution)	0.014	0.002	0.011	-0.008	-0.006
	[0.023]	[0.031]	[0.023]	[0.033]	[0.032]
National Percentile Rank	0.384	0.007	0.256	-0.323	-0.273
	[0.665]	[0.957]	[0.690]	[1.013]	[0.971]
GPA in Non-Math Courses 2013-2014, N = 2,312					
Non-math GPA, SY2013-14	0.072*	0.067*	0.069*	0.068*	0.067*
	[0.026]	[0.024]	[0.025]	[0.026]	[0.026]
Non-math GPA, z-score in control group SD's	0.069*	0.064*	0.067*	0.066*	0.064*
	[0.025]	[0.023]	[0.024]	[0.025]	[0.025]
Non-Math Courses Failed 2013-2014, N = 2,312					
All non-math course failures, SY2013-14	-0.347**	-0.334**	-0.348**	-0.353***	-0.349***
	[0.102]	[0.077]	[0.097]	[0.078]	[0.076]
All core non-math course failures, SY2013-14	-0.165*	-0.163*	-0.170*	-0.182**	-0.178*
	[0.065]	[0.060]	[0.059]	[0.057]	[0.059]
Outcome Domain: Behavior					
Discipline Incidents SY13-14, N = 2,575	0.066	0.086	0.091	0.088	0.082
	[0.145]	[0.143]	[0.144]	[0.151]	[0.141]
Days Absent SY13-14, N = 2,575	0.543	0.699	0.518	0.651	0.592
	[0.922]	[0.767]	[0.936]	[0.956]	[0.905]
Out-of-School Suspension Days, N = 2,575	0.178	0.219	0.209	0.215	0.211
	[0.216]	[0.208]	[0.209]	[0.218]	[0.195]
Outcome Domain: Crime					
Arrests for Violent Crimes Aug. 26, 2013--Jul. 1, 2014, N = 2,718	-0.016	-0.014	-0.016	-0.014	-0.014
	[0.010]	[0.010]	[0.009]	[0.010]	[0.010]
Arrests for Property Crimes Aug. 26, 2013--Jul. 1, 2014, N = 2,718	-0.004	-0.004	-0.006	-0.004	-0.005
	[0.011]	[0.011]	[0.011]	[0.010]	[0.012]
Arrests for Drug Crimes Aug. 26, 2013--Jul. 1, 2014, N = 2,718	0.013	0.013	0.012	0.013	0.012
	[0.014]	[0.016]	[0.016]	[0.015]	[0.017]
Arrests for Other Crimes Aug. 26, 2013--Jul. 1, 2014, N = 2,718	-0.03	-0.03	-0.033	-0.027	-0.032
	[0.019]	[0.021]	[0.020]	[0.019]	[0.023]

+ p < 0.1; * p < 0.05; ** p < 0.01; *** p < 0.001

Model information: randomization block number fixed effects included in ALL models. Standard errors clustered within schools. All students with non-missing outcome data for each measure are included in the models.

Baseline Demographics: indicator for age 14, indicator for age 15, indicator for 10th grade, indicator for free lunch recipient, indicator for learning disability, indicator for black, indicator for hispanic,

Baseline 2012-13 academic measures: Overall GPA , number of absences, and Explore/Plan reading and math scaled score (standardized to control group). For each of these variables, missing observations were recoded as 0's and indicators such as "missing 2012-13 GPA" were created and included in the models.

Crime outcomes include the following additional baseline measures: number of prior violent arrests , number of prior property arrests, number of prior drug arrests, number of prior arrests which are not violent, drug, or property related.

Table 7. Match ITT and TOT effect for ISR Test

ISR Outcomes (N=641)	Outcome Type	Control mean (original)	Control SD (original)	Control mean (recode)	ITT *b*	ITT *se*	TOT *b*	TOT *se*
Social								
Youth reports having no close friends	Dichotomous	NA`	NA	0.15	0.007	0.085	0.011	0.134
To what degree do you agree that your friends think it is important to:								
"Attend classes regularly"	z-score	3.72	0.52	0	-0.018	0.082	-0.027	0.127
"Get good grades"	z-score	3.7	0.58	0	-0.041	0.079	-0.063	0.123
"Study"	z-score	3.37	0.77	0	-0.188*	0.087	-0.291*	0.135
"Continue education"	z-score	3.71	0.55	0	0.065	0.083	0.101	0.130
Over the current school year, have you:								
"Stopped hanging around with anyone because you thought that spending time with them was likely to put you in a situdation that could lead to trouble"	Dichotomous	NA	NA	0.51	0.036	0.043	0.056	0.068
"Started hanging around with anyone because you thought spending time with them was likely to keep you out of situations that could lead to trouble."	Dichotomous	NA	NA	0.61	0.009	0.043	0.014	0.067
Academic								
"Thinking about your school, how much do you agree with the statement: Disruptions by other students get in the way of my learning"	z-score	2.23	0.84	0	-0.070	0.090	-0.110	0.141
"Overall about how much total time do you spend on homework each week, both in and out of school"	z-score	1.97	0.79	0	-0.029	0.090	-0.046	0.142
"When homework is assigned, how much of it do you usually complete?"	z-score	3.99	0.97	0	0.079	0.085	0.125	0.133
"As things stand now, how far in school do you think you will get"	Dichotomous	NA	NA	0.91	0.023	0.022	0.036	0.035
"As things stand now, how far in school do you want to go"	Dichotomous	NA	NA	0.93	0.018	0.021	0.028	0.033
"How important are good grades to you?"	z-score	3.49	0.642	0	0.145+	0.080	0.227+	0.125
"This school year, how often did you feel safe at your school"	z-score	3.23	0.76	0	0.058	0.089	0.090	0.140
How true would you say the following statements are:								
"I like math"	z-score	2.74	1.06	0	0.142+	0.084	0.223+	0.133
"I get good grades in math"	z-score	2.77	0.95	0	0.233**	0.076	0.364**	0.121
"In general, I like school"	z-score	2.79	0.9	0	-0.070	0.085	-0.109	0.132
"I like reading"	z-score	2.52	1	0	-0.088	0.080	-0.138	0.125
"I get good grades in reading	z-score	2.93	0.88	0	-0.001	0.078	-0.002	0.122
Relationships with Adults								
"How many adults do you have in your life who you feel comfortable talking to about personal problems"	integer	4.2	4.12	NA	0.018	0.314	0.028	0.493
"How many aults do you have in your life who care a lot about how you turn out and who will help you if you get into trouble"	integer	7.21	7.4	NA	0.348	0.566	0.546	0.885
"Do you talk to an adult at school when you need help with your school work?"	Dichotomous	NA	NA	0.37	-0.007	0.040	-.012	0.063

p < 0.1 +
p < 0.05 *
p < 0.01 **
p < 0.001 ***

Coding: See 'Question Codebook' for details about specific survey questions and response coding+.

Means and SDs are unweighted and z-scores were calculated using sample means and sd's.

Regression estimates are weighted and standared errors were calculated using Taylor Series linearization.