Implementation Matters: Experimental Evidence from Comprehensive High School Supports

Thomas Dee, Chloe R. Gibbs, Rebecca Margolit-Chan, Lillian Morrell, and James Sullivan *

September 25, 2025

Preliminary Draft

Abstract

High school completion and successful transitions to post-secondary schooling or the workforce are vital steps for economic well-being in the United States. While high school graduation rates have steadily increased over the last three decades, disparities persist across schools and districts and between youth from high- and low-income households with important ramifications for longer-run economic outcomes. To address these gaps, schools have increasingly adopted or partnered with comprehensive support programs providing wraparound services. This randomized controlled trial, deployed in two urban, disadvantaged school districts, tests the effectiveness of a mature comprehensive support program targeted to high school students on an unusually large scale (n=2,528). While overall effects on graduation rates were small and statistically insignificant, exploratory analyses reveal that this null result masks notable variation by site and over time. These findings highlight the challenges of scaling interventions and raise concerns about the generalizability of results from small-scale, single-site evaluations.

*Dee: Stanford University & NBER (email: tdee@stanford.edu); Gibbs: University of Notre Dame, W.E. Upjohn Institute for Employment Research & NBER (email: chloe.gibbs@nd.edu); Margolit-Chan: Cornell University (email: rpm248@cornell.edu); Morrell: University of Notre Dame (email: lmorrell@nd.edu); Sullivan: University of Notre Dame (email: jsulliv4@nd.edu). This research was supported by the University of Notre Dame's Wilson Sheehan Lab for Economic Opportunities (LEO) and MIT's Abdul Latif Jameel Poverty Action Lab (J-PAL). We are grateful to the two participating school districts and the state department of education for their cooperation and assistance with data collection. We also appreciate excellent project management and research assistance from Brendan Perry, Grace Ortuzar, Emily Merola, Anthony Tatarka, and Ethan Vonhedemann. Any remaining errors are the responsibility of the authors.

I Introduction

Dropping out of high school is often characterized as a form of "economic death sentence" (Cullen et al., 2013) as the shift towards a knowledge-based economy has made low-education jobs scarcer and less financially viable. This concern reflects a long-standing body of causal evidence, which shows that obtaining a high school diploma confers substantial improvements in lifetime earnings, employment, and health (Card, 1999). In recognition of its importance, high school completion, along with college and career readiness, is now embedded as a key performance indicator in the state accountability systems developed under the Every Student Succeeds Act (ESSA), the 2015 reauthorization of the Elementary and Secondary Education Act.

Although high school graduation rates have increased dramatically in the United States over the last 50 years, concentrations of high school non-completion exist, especially in under-resourced schools and among disadvantaged students. These pockets of persistently low graduation rates have motivated policy-focused research on programs designed to support high school completion among educationally vulnerable students. Structured reviews of this literature sponsored by the U.S. Department of Education's Institute of Education Sciences (IES) underscore the promise of programs that feature academic and social-emotional supports, adult mentors who provide individualized attention to at-risk students, and connections between schoolwork and post-secondary college and career opportunities (Dynarski et al., 2008; Rumberger et al., 2017).

In particular, initiatives that integrate such "wraparound" supports and services have attracted considerable attention. For example, the Check and Connect program features adult mentors who monitor student attendance and performance and connect students to aligned social services and academic supports. Early random-assignment evaluations (Sinclair et al. 1998, 2005) found that this approach was effective in reducing dropping out when fielded among small populations of students with special learning needs (n=94 and n=147, respectively). However, field-experimental studies situated in somewhat larger and more general student populations have yielded at best mixed results. For example, Maynard et al. (2014) report gains in academic grades, though not attendance, among middle and high school students (n=260). In a study of 9th graders (n=553), Heppen et al. (2018) find that these services had no detectable effects on high school completion or other measures of student engagement and progress. Relatedly, experimental results from a large-scale implementa-

tion of Check and Connect among elementary and middle-school students in Chicago Public Schools (i.e., n=9,489) found no effects on multiple indicators of academic success (e.g., test scores and grades) but did find attendance gains limited to the middle school level (Guryan et al., 2020).

This pattern of results is consistent with an established inverse relationship between impact and scale in the broader program-evaluation literature. For example, Kraft (2020) examined 1,942 impact estimates reported in randomized control trials of educational interventions and found that the average effect sizes (ES) reported in the smallest studies (i.e., ≤ 100 students, ES=0.30; 101-200 students, ES=0.15) are *3 to 6 times* larger than in the largest studies (i.e., $\geq 2,000$ students, ES=0.05). This stark pattern may simply reflect a selection bias in which higher-impact but expensive interventions are more commonly evaluated in smaller-scale trials. However, these comparative results can also be understood as an indicator of the considerable and well-known difficulty of maintaining a program's implementation fidelity at scale.

This study contributes to our understanding of these issues through the large-scale experimental evaluation of a mature program designed to support the success of educationally vulnerable students with comprehensive high school supports (CHSS). This preregistered randomized control trial (RCT) is situated in two large urban school districts and examines high school outcomes among three district-year cohorts of 9th grade students (n=2,381). The CHSS intervention has multiple features, including mentoring and case management provided by professional staff, coordination of various academic supports (e.g., tutoring and homework assistance), family engagement services, and assistance with college and career readiness (e.g., support for post-secondary access, resume development, and interview preparation).

Overall, we find that random assignment to this mature program had small and statistically insignificant effects on the primary outcome of interest, high school graduation. However, the unusual scale of this RCT and the richness of the available data allow us to explore key aspects of the program implementation and any underlying heterogeneity in impact. We find that the overall null finding masks extraordinary heterogeneity in program implementation and impact across district-year cohorts. In particular, in one of the two participating school districts, assignment to the program had a positive impact on high school graduation in one cohort but a negative impact on a subsequent cohort, entering just one year later.

We examine these findings within a well-known framework that characterizes treatment

heterogeneity with respect to the "Three C's": client characteristics, program context, and the treatment contrast (Weiss et al., 2014). While student characteristics appear largely consistent across the three site-years, we observe some notable contextual differences. Two of the three district-year cohorts were completing their intended senior year during the first full school year of the COVID-19 pandemic, 2020–2021, while the remaining cohort was in its third year. The study districts also adapted their graduation policies differentially for these cohorts.

However, the most stark differences involved the treatment contrast for each district-year cohort, much of which occurred *before* the onset of the COVID-19 pandemic. In particular, we find substantial differences in the timing, composition, and intensity of the treatment services students received. In the site-year with positive impacts, program takeup occurred immediately when students began their grade-9 year and it emphasized higher participation in tutoring and academic enrichment relative to other cohorts. However, we lack comparable data on the experiences of students not assigned to the program, which limits our ability to fully characterize the treatment contrast.

We view this study as making several distinct contributions. Most obviously, this study contributes to the comparatively large experimental literature on the design and impact of dropout-prevention programs targeted to high school students. In particular, this study provides evidence on the impact of comprehensive supports when evaluated on a uniquely large scale. The null results suggest that the enthusiasm for providing high school students with a multifaceted set of supports is misplaced if the challenges of consistently implementing these supports in a high-quality manner are not addressed. Alternatively, to the extent the high-fidelity implementation of wraparound services cannot feasibly scale, these results argue for the policy appeal of other program designs. These include cost-effective, "light touch" interventions that are scalable (e.g., text messaging about attendance) and initiatives that have shown promise on a larger scale with a more singular approach such as early college access (Berger et al., 2013; Edmunds et al., 2015), cognitive behavioral therapy (Heller et al., 2013, 2015), and career academies (Kemper, 2001).

This study also makes two broader contributions to the program-evaluation literature. First, this study has implications for how we understand the general evidence of an inverse relationship observed between program impact and scale. Specifically, these study results (i.e., a null result for a more complex and expensive intervention evaluated on a comparatively large scale) are *not* consistent with the argument that the effect-size/scale relationship

merely reflects selection biases. However, the evidence presented in this study is consistent with the view that this relationship reflects the difficulty of consistently maintaining the quality of more complex programs on a larger scale. The broader experimental literature on wraparound services for dropout prevention also support this interpretation in that the most encouraging evidence of impact appear in the smallest studies.

Second, we view our results as a cautionary tale about the risks of using evaluation evidence in an overly simplistic or credulous manner. For example, the study results presented here illustrate how focusing on a single site in a single year could yield highly divergent conclusions about a program's impact relative to another similar site or even to the same site one year later. This issue is especially relevant given that long-lived research consensuses about particular interventions are often built on the results of relatively few RCTs that rely on data from highly selected sites, small samples, and few time periods. Relatedly, these results also constitute a compelling reminder of the importance of using evaluation evidence not just summatively but formatively. Specifically, this study's evidence of treatment heterogeneity related to implementation fidelity underscore how rigorous, localized evaluation can be a powerful tool in support of the "continuous improvement" of conceptually compelling programs (Dee, 2024).

II The Intervention

We study the impact of a comprehensive, school-based support program designed to improve high school graduation rates among students at elevated risk of dropout, which we refer to as Comprehensive High School Supports (CHSS). The program was launched in the 1980s and has since grown into a widely-deployed and nationally recognized model, currently serving over 4,000 students each year across multiple school districts in two states. Participants are typically enrolled at the start of 9th grade and remain engaged through graduation.

At the core of the CHSS model is a full-time, professional case manager assigned to each student. These staff members deliver long-term, individualized mentoring and case management, connecting students to a coordinated web of academic, social, and emotional supports. The CHSS model is guided by an evidence-informed approach that emphasizes relationships, consistency, and responsiveness to student needs.

CHSS services span five key domains:

1. Mentoring and Case Management: Regular one-on-one meetings with case managers, who support academic progress, behavioral goals, and personal development.

2. Academic Support: Tutoring, homework help, and test preparation, particularly for state-mandated graduation exams.

3. Family Engagement: Outreach to caregivers through monthly check-ins, connecting families to community resources and ensuring continuity of support at home.

4. After-School Enrichment: Activities designed to build life skills, foster positive peer relationships, and promote school engagement.

5. College and Career Readiness: Exposure to post-secondary education pathways and job readiness training, including resume development, interview preparation, and paid internships.

These supports are embedded within a broader model, which integrates school, family, and community partnerships to provide a coordinated system of care for each student. Case managers coordinate closely with teachers, school counselors, and external service providers to align interventions and address both academic and non-academic barriers to success.

During the COVID-19 pandemic, CHSS adapted core services for remote delivery. Mentoring and academic support were provided virtually, and the frequency of contact with students and families significantly increased. The program also took on a stabilizing role by distributing learning materials, technology access resources, and basic necessities such as food.

The program operates in districts that primarily serve socioeconomically disadvantaged youth. These districts collectively enroll approximately 40,000 students and invest around $22,000 per student annually.[1] Despite this investment, the average graduation rate in the two districts in 2021 was 70 percent, well below the national average of 86 percent (National Center for Education Statistics, 2025).

Understanding the context of each district further underscores the need for targeted intervention. District A serves over 20,000 students, with half of families living below the poverty line, more than 60 percent receiving Supplemental Nutrition Assistance Program (SNAP) benefits, and nearly 70 percent of students from single-mother households. Over 10 percent

[1]Descriptive statistics come from publicly available district data.

of students have documented disabilities. District B serves a similarly sized population with comparable demographic and economic profiles.[2]

These conditions underscore the relevance of CHSS as an intervention tailored to the multifaceted barriers students face in completing high school. This evaluation examines the impact of the CHSS model on academic outcomes.

III Study Design and Enrollment

To determine the effect of this comprehensive program of student and family supports on academic progress and graduation, we conducted a randomized controlled trial (RCT) that included ninth grade students entering high school from two public school districts, referred to as "District A" and "District B." Both districts are located in urban areas and serve many low-income students who are at high risk of not completing high school. For District A, we enrolled one cohort during the 2017–2018 school year, while for District B, we enrolled two cohorts: 2017–2018 and 2018–2019.

3.1 Constructing the Study Sample and Implementing Randomization

To construct our study sample, we relied on student-level administrative data from the two school districts. At the beginning of the school years during which we were enrolling students, the districts provided the research team with de-identified lists of eligible incoming ninth-grade students at the 16 schools that offered the intervention. These lists included all ninth graders who met at least two out of six predetermined risk factors associated with dropout rates and had accrued no more than 17 days of suspension. The suspension criterion is enforced because extensive suspension days are correlated with unresolved behavioral issues that the program may not be equipped to address. The six factors include:

1. Low socioeconomic status, defined as eligibility for free or reduced-price lunch
2. Poor school attendance, indicated by attendance rates falling between 71 percent and 85 percent

[2]Student demographic data are drawn from the NCES Common Core of Data.

3. Over-age for grade, signifying being behind by 1 or 2 grades
4. Failure in two or more English or math courses
5. Subpar standardized test scores, categorized as Level 1 or 2 on the state assessment or scoring below the 60th percentile on any nationally normed standardized test
6. Having multiple school suspensions

These eligibility criteria were designed to target students who would be the most likely to benefit from this unique intervention: those who are capable of success but in danger of dropping out of school.

The number of students satisfying these eligibility criteria included 1,005 students from District A, 739 from District B in 2017–2018, and 784 from District B in 2018–2019, totaling 2,528 students. The program could serve about 250 students in District A and 200 in District B in any given year, with the number of available slots varying by school. To determine which students from the pool of eligible students to invite to participate in the program, the research team constructed randomly ordered lists of students stratified by school, cohort, gender, and "risk factor" category (either two risk factors or more than two). The research team sent these randomly ordered lists to the program's senior administrators, who linked the de-identified lists to personal identifying information. Program staff at each school then received the names of students to recruit for program participation, provided in the order generated by the research team. Program staff would reach out to these students in the order that names were received (student names were provided in small batches of about five or so). When students could not be contacted after several attempts, or students chose not to participate, program staff would move on to the next student on the list, continuing this process until capacity was reached. Students from the lists who were not invited to participate in the program continued to have access to the other usual services offered by the school district. Each district provided a range of other support services, including school counseling, access to social worker and psychologists, academic tutoring, extended learning time, and referrals to community-based resources. Students could also participate in social-emotional learning (SEL) activities and career and technical education (CTE) programs, which provided industry-aligned training and skill development. However, no program providing coordinated, individualized, and sustained support—comparable to what students receive with CHSS engagement—was available to students in the control group.

3.1.1 Assigning Students to the Treatment Group: Initial Offer vs. Ever Offer

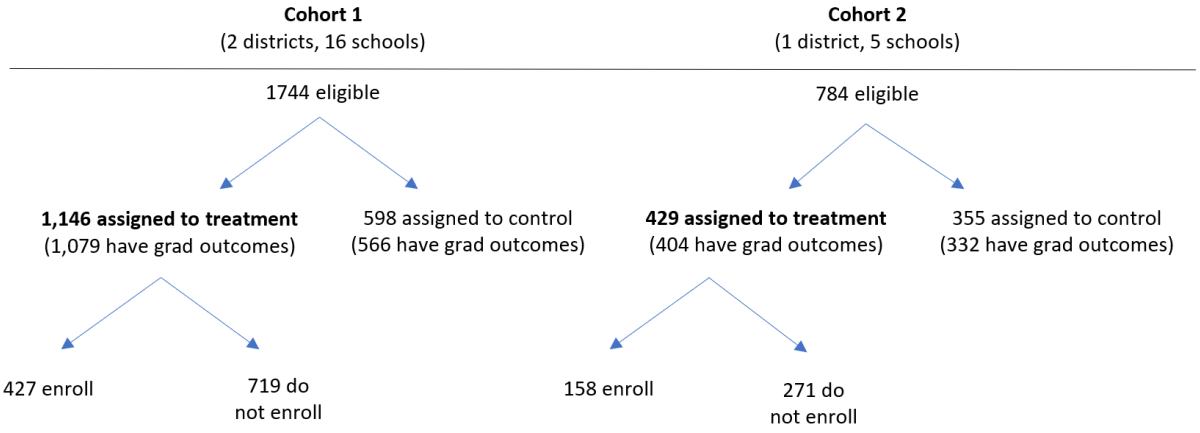
Because of endogeneity of program takeup in a randomized waitlist setting (de Chaisemartin and Behaghel, 2020), we consider two definitions of "treatment offer." The first approach, "ever offer," assigns the treatment offer to any student whose name was released to the program staff for recruitment, with unreleased names constituting the control group. The second approach, "initial offer," assigns 60 percent of each school's randomly ordered list, in numeric order, to the treatment group and the remaining 40 percent to the control group. This 60 percent-40 percent breakdown was based on projections of program takeup based on information from program staff. In this latter approach, some students are assigned to the treatment group even though they are never offered treatment, because the program reached capacity before working through 60 percent of the waitlist. This approach avoids bias from any individual student's probability of receiving an offer being conditional on the takeup rate among preceding students on the list. For all of our main analyses, we define treatment status based on the "ever offer" categorization, but in our robustness section, we present results using the "initial offer" definition, showing that how we define treatment status does not substantively affect our results.

Figure 1 displays the number of program-eligible students in each cohort and their assignment to treatment or control groups based on the "ever offer" approach. A total of 2,528 students were eligible for the program, 1,744 from the 2017 cohort and 784 from the 2018 cohort. Of these, 1,575 were offered treatment, leaving 953 in the control group. However, our primary analysis that examines the effect of the program on high school graduation relies on a large subset of this study sample, because graduation status is missing in school district records for 5.8 percent of the students. We examined whether treatment status was related to the likelihood that a student's graduation status was missing. In general, we do not find noticeable differences in missingness by treatment status (see appendix Tables A.1 and A.2 for additional detail.).[3] We observe graduation status for 1,645 students from the

[3]Specifically, we define an outcome, Missing Grad Outcome, as an indicator of whether a student had any of the following statuses that did not culminate in a graduation or dropout outcome: transferred to a high school equivalency program, other transfers, and unknown statuses. As we do not observe whether a student who transferred to a high school equivalency program obtained any certificate or diploma, we include this outcome as a missing outcome. Other transfer statuses include those who transferred out of district, out of state, and out of country, and are unobservable in our dataset. Students with unknown statuses are students for whom the district did not receive any documentation. While we find little difference in missingness by treatment status, the results in Appendix Table A.2 do show that students who were assigned to the treatment group in the second cohort of District B were approximately 4 percentage points (almost 50 percent) more likely to transfer to a high school equivalency program compared to those in the control

2017 cohort and 736 from the 2018 cohort. Therefore, our analysis sample includes 2,381 students with observed graduation status.

Figure 1: Treatment Assignment



Note: This figure reports the number of students in the study by treatment status. The figure also reports the number of students who enrolled in the program among those assigned, and the number of students for whom we observe graduation outcomes. Ever offer is used as the treatment definition.

3.1.2 Program Takeup and Participation

As shown near the bottom of Figure 1, many students assigned to receive the program chose not to enroll (about two-thirds). For instance, in the 2017 cohort, 1,146 students were assigned to the treatment group, but only 427 ultimately enrolled in the program, as indicated by having a program start date. Appendix Figure A.1 shows that the takeup rate was similar across school districts and cohorts. The overall takeup rate was 37 percent, and all of the school district-cohort samples had takeup rates between 30 and 39 percent. Additional analyses show that program participation was higher among eligible students with only two risk factors.

group. However, we find that this effect is driven by those who did not take up the program.

3.2 Characteristics of the Study Sample and Balance by Treatment Status

Table 1 reports the baseline characteristics for the 2,528 students in the full study sample as well as the 2,381 students in our analysis sample that had observable graduation outcomes. In the first three columns, we report means for the full sample, the control group, and the treatment group, respectively. In the fourth column, we report the regression-adjusted difference between these means, including as covariates: whether the student is female and whether the student had over two risk factors, as well as school at baseline and cohort fixed effects.

The results show that the baseline characteristics of the treatment and control groups for the full sample are very similar, with little statistically significant difference between them. This indicates that overall, the treatment and control groups were balanced on observable characteristics at the start of the study and that the random-order waitlist successfully produced comparable groups of students. We find a similar balance in baseline characteristics for the analytic sample, which includes students with observable graduation status by the end of our study period. We find a statistically significant difference in age by treatment status, but this difference is very small (about half of a month).

These baseline characteristics also emphasize how the CHSS program targets extremely disadvantaged students. Nearly all of the students live in poor households; half have more than two risk factors (only two are required for eligibility); for a quarter of them, English is not their native language; and the predicted graduation rate (based observable baseline characteristics) is 73 percent, well below the national average.

Table 1: Baseline Balance, Full and Analytic Samples

	All	Control	Treatment	Adj. Difference
	Means	Means	Means	Treatment - Control
	(a) Full Study Sample			
White	0.161	0.173	0.154	-0.013
				(0.015)
Black	0.530	0.518	0.537	0.016
				(0.020)
Hispanic	0.207	0.200	0.212	0.00035
				(0.016)
English Second Language	0.266	0.276	0.259	-0.017
				(0.018)
In Poverty	0.934	0.931	0.936	0.0045
				(0.010)
Enrollment Age (Years)	14.10	14.11	14.09	-0.024
				(0.025)
Predicted Graduation Rate	0.728	0.729	0.727	-0.0013
				(0.0032)
Female	0.476	0.472	0.479	0.0039
				(0.020)
Over Two Risk Factors	0.531	0.529	0.532	0.0064
				(0.020)
Total Obs.	2,528	953	1,575	2,528
	(b) Analytic Sample			
White	0.159	0.175	0.150	-0.019
				(0.016)
Black	0.540	0.530	0.547	0.015
				(0.021)
Hispanic	0.204	0.196	0.209	0.0020
				(0.017)
English Second Language	0.250	0.262	0.242	-0.020
				(0.018)
In Poverty	0.934	0.930	0.936	0.0043
				(0.010)
Enrollment Age (Years)	14.09	14.11	14.08	-0.032
				(0.026)
Predicted Graduation Rate	0.729	0.730	0.728	-0.00014
				(0.0033)
Female	0.480	0.473	0.485	0.0093
				(0.021)
Over Two Risk Factors	0.531	0.524	0.535	0.016
				(0.021)
Total Obs.	2,381	898	1,483	2,381

Notes: *** p<0.01, ** p<0.05, * p<0.10. Statistics are taken from the data provided by the school districts. The difference column is regression-adjusted by whether the student is female and whether the student had over two risk factors. Predicted graduation rate is a predicted value of the graduation rate based on the other covariates in the table. The number of observations for enrollment age and predicted graduation is two less for the full sample due to missing information. The full sample include all students who were randomized for the study. The analytic sample restricts the full sample for those whom we observe main outcomes. Ever offer is used as the treatment definition.

IV Data and Methods

4.1 School District and State Education Department Data

Our empirical analysis leverages student-level administrative records from the two public school districts where the study was implemented. Before randomization, the school districts provided the research team with files containing baseline characteristics of entering 9th grade students, which were used to determine eligibility for the study. Following each academic year, the districts continued to send updated files with additional characteristics and academic outcomes, allowing us to establish a student-by-year panel for assessing both academic and non-academic outcomes.

We linked these student records to data from the National Student Clearinghouse (NSC), enabling us to observe post-secondary education outcomes, such as enrollment in two- or four-year colleges or vocational programs, the name of the institution, the number of semesters enrolled, whether enrollment was full or part-time, and graduation status[4].

The administrative data provided by the school districts include key demographic indicators such as birth date, gender, race, ethnicity, language spoken at home, eligibility for free and reduced-price lunch, English language learner status, and the total number of risk factors. In District A, low socioeconomic status is defined as household income below the federal poverty line, while in District B, it is indicated by eligibility for free or reduced-price lunch. District B also provides information on whether a student has a documented disability, defined as having either a 504 Plan or an Individualized Education Program (IEP).

Our primary outcomes, specified in advance in our pre-analysis plan, are two measures of high school persistence and completion. Successful high school completion is captured by whether a student earned a high school diploma within four or five years, with graduation determined by receipt of a diploma by June 30 of the relevant year. These outcomes are constructed using longitudinal student-level data from the participating school districts, supplemented with statewide administrative records on high school enrollment and exit. If a student is identified as graduated in either district or state data, we classify them as having completed high school. If the district data are missing or incomplete, we rely on state records

[4]The National Student Clearinghouse is a nonprofit organization that works with more than 3,600 post-secondary institutions to provide data on student enrollment and degree completion. The NSC-participating colleges enroll 98 percent of all post-secondary students in the US. (https://www.studentclearinghouse.org/about/).

to determine graduation status.[5]

Additionally, we examine several closely related outcomes that shed light on students' academic progress and pathways through high school. These include whether a student dropped out of high school, whether they were "on-track to graduate" (as specified in pre-registration), and the individual components of that index in later high school years.

The "on-track to graduate" index is a composite index designed to capture early indicators of academic progress toward high school completion, constructed based on a review of related literature (Allensworth and Easton, 2005; Kemple and Stephenson, 2013). The index incorporates five components, each coded as a binary indicator based on student-level administrative records. A student is considered "on-track" in a given year if they meet all five criteria: (1) the student did not drop out during their time in high school; (2) the student was not chronically absent, defined as attending fewer than 90 percent of enrolled instructional days, conditional on being enrolled for at least ten days and present for at least one; (3) the student met a minimum credit accumulation threshold of 5 credits during their first year, and 10 credits in their second year if enrolled in District A; (4) the student participated in required state assessments; and (5) the student earned at least one passing score on a state exam. The index draws on prior research identifying early academic benchmarks as strong predictors of eventual graduation and serves as a summary measure of academic progress during high school (Kemple and Stephenson, 2013; Allensworth, 2005). The composite index, and its component parts, allow us to examine the potential pathways through which the intervention did and did not affect high school completion.

4.2 Provider Data on Participation and Activity Engagement

The implementing partner provided detailed administrative records on student participation in the CHSS program. These records included information on the duration of each student's engagement, the frequency of contact with their assigned case manager, and logs of participation in specific program activities. For each activity, the data captured its type, frequency, and duration. The provider also shared staff-level data, including unique identifiers

[5]Students are classified as having graduated only if they earned a diploma recognized by the state as fulfilling high school graduation requirements. Students who earned a High School Equivalency (HSE) certificate or a Commencement Credential are not considered graduates. Students whose final status indicates dropout, long-term absence, aging out without a diploma, or withdrawal without documentation are also classified as not graduated. Students recorded as transferring out of the district are considered graduates only if diploma receipt is observed in state data.

for case managers and indicators of staff turnover during the study period.

4.3 Analytic Approach: ITT and TOT

Given the random assignment of students to treatment and control groups, we estimate the program's impact using an intent-to-treat (ITT) model, which captures the difference in outcomes between students assigned to the treatment and control groups, regardless of actual program participation. In particular, we estimate:

$$Y_{ics} = \beta_0 + T_{ics}\beta_1 + X_{ics}\beta_2 + \theta_c + \lambda_s + \epsilon_{ics} \tag{1}$$

Where Y_{ics} is an indicator for one of our outcome measures for student i. T_{ics} is an indicator for whether student i from the randomly ordered waitlist in cohort c and school s was ever offered to enroll in the CHSS program.[6] The vector X_{ics} includes a set of person-level characteristics collected at baseline, including gender, race/ethnicity (white, Black, or Hispanic), whether English is a second language, poverty status, indicators for age at enrollment (15 or 16+), and whether they have more than two risk factors.[7] θ_c and λ_s are cohort and school fixed effects respectively, and ϵ_{ics} is an error term.

To estimate the effect of the program on those who actually participated (treatment-on-the-treated, TOT), we estimate an instrumental variable (IV) model, where treatment assignment serves as an instrument for program participation. The TOT model is specified as:

$$Y_{ics} = \gamma_0 + \hat{P}_{ics}\gamma_1 + X_{ics}\gamma_2 + \theta_c + \lambda_s + \eta_{ics} \tag{2}$$

where $\hat{P}_i$ is a prediction for participation in the program, which is generated from a first stage equation where we regress program participation on the treatment assignment indicator as well as cohort and school fixed effects and observable characteristics. Program participation (or takeup) is defined as completing the program enrollment process. While the ITT reflects the effect of being offered the program, the TOT is an estimate of the effect of participating

[6] As noted earlier, for all of our main analyses, we define treatment status based on the "ever offer" categorization, but in our robustness section, we present results using the "initial offer" definition, showing that that the choice of treatment definition does not substantively affect our results.

[7] Specifically, for each covariate with missing data, we create a binary indicator for whether the value is missing and set the covariate itself to a fixed constant for those cases. We then include both the covariate and its missingness indicator in the regression. This approach allows us to retain all students in the analysis sample while accounting for missing data in a way that avoids biasing coefficient estimates.

in the program (Angrist and Pischke, 2009).

V Results

In this section we report the treatment effects for our primary outcome, high school graduation, for our full sample as well as separately by district and cohort. We also examine heterogeneity in treatment effects across several demographic groups. We pre-registered subgroup analyses by gender, race and ethnicity, age, predicted likelihood of graduation, and baseline risk factors as confirmatory tests of treatment effect heterogeneity. Analyses disaggregated by district and cohort, as well as those examining other student characteristics such as special education and English language learner status were not pre-specified and are therefore exploratory. We additionally show effects on a pre-registered "on-track to graduate" index and on intermediate high school outcomes and post-secondary enrollment outcomes, both of which were defined as exploratory in the pre-analysis plan.

5.1 High School Completion

We estimate treatment effects on four- and five-year high school graduation rates for the full sample and across three separate district-cohort groups: one cohort that entered high school in the 2017–18 school year in District A, and two entering cohorts, in 2017–18 and 2018–19, in District B. ITT estimates, which are reported in Table 2 can be interpreted as the average effect of the program offer. We use the "ever offer" definition of treatment and present two approaches to statistical inference, with heteroskedasticity-robust standard errors and standard errors clustered at the level of randomization (school-risk factor-gender-cohort). We pre-specified clustered standard errors in the pre-analysis plan to align with our randomization design. However, in line with recent guidance on best research practice, we also report heteroskedasticity-robust standard errors for completeness (Abadie et al., 2023).

The pooled estimates presented in Column (1) suggest that the program has no effect on four- or five-year graduation. For our primary outcome, four-year graduation, the point estimate is very small (-0.012) and not statistically different from zero. We can rule out positive effects greater than 2.3 percentage points, or a 3.2 percent increase relative to the mean for the control group. This null finding for the pooled sample, however, masks important variation in the treatment effect across districts and cohorts. For district A

(column 2) we also see little evidence of a treatment effect. For district B (columns 3 and 4), the treatment effects differ sharply across the two cohorts. Students in the 2017 cohort who were offered the program were about six percentage points more likely to graduate, which is a large effect compared to the control group mean of 73 percent, although this effect is only marginally significant using the clustered standard errors. In contrast, for the 2018 cohort, treatment group students were significantly less likely to graduate in four (-8.4 percentage points) or five years (-7.1 percentage points), and the effect for four-year graduation is statistically significant.

Appendix Table A.3 presents the TOT results for the same groups as we report in Table 2. In addition to the point estimates, we report control complier means.[8] Again, we see no evidence of a treatment effect for the full sample, but we do see noticeable differences in the treatment effects across the two cohorts in District B. For the first cohort (Columns 6), the program led to a 14.5 percentage-point increase (p-value: 0.120) in the likelihood of graduating in four years years and a 15.2 percentage point increase (p-value: 0.099) after five years. In contrast, for the second cohort the TOT estimate is negative and significant for both four- and five-year graduation, and these effects are statistically significant. Our exploratory analysis suggests that the reduced likelihood of graduation for 2018 cohort was driven by non-compliers, those assigned to treatment who did not take up the program. Baseline characteristics for this cohort suggest that this was a particularly at-risk group. As we discuss further below, for this cohort the program may have struggled to reach vulnerable students in the critical early period of their high school trajectories due to staff turnover, implementation delays, and other challenges.

[8]The control complier mean, which reflects the average outcome for those in the control group who would have taken up the treatment if offered, is calculated as the mean of the outcome for treatment compliers minus the TOT treatment effect estimate.

Table 2: ITT Effect of Program on 4- and 5-Year Graduation, Pooled and By District

	Pooled	District A	District B	
	C17 & C18	C17	C17	C18
	(a) 4-Year Graduation			
	(1)	(2)	(3)	(4)
Treatment	**-0.012**	0.014	0.057	-0.084
	(0.018)	(0.027)	(0.033*)	(0.032***)
	[0.019]	[0.026]	[0.035]	[0.031**]
Control Mean	**0.7294**	0.7143	0.7257	0.7470
N	**2,381**	952	693	736
	(b) 5-Year Graduation			
	(1)	(2)	(3)	(4)
Treatment	**-0.004**	0.021	0.060	-0.071
	(0.017)	(0.026)	(0.033*)	(0.031**)
	[0.020]	[0.023]	[0.035]	[0.040*]
Control Mean	**0.7572**	0.7508	0.7342	0.7801
N	**2,381**	952	693	736

Notes: Bolded analyses were pre-registered. *** p<0.01 ** p<0.05 * p<0.1. All estimates include school-level and cohort fixed effects. ( ): robust unclustered standard errors, [ ]: standard errors clustered by school/risk factor/gender/cohort. This table uses ever offer as the treatment definition. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. These results are for our analytic sample, which includes the 2,381 students with observable graduation outcomes.

5.2 Robustness

We conduct several additional analyses to probe the robustness of our main findings. First, we employ alternative specifications of our main regression model to examine the sensitivity of our estimates to different model specifications. Specifically, we test our alternative definition of treatment: "initial offer" rather than the “ever offer” definition of treatment used for

our main analyses." The estimates using "initial offer" (Appendix Table A.4) are very similar to the estimates using "ever offer" (Table 2) both in terms of the size and precision of the estimates. This is not surprising given that the fraction of the sample assigned to treatment using the "initial offer" approach (66 percent) is very close to the fraction assigned treatment using the "ever offer" approach (62 percent).

In addition, our preferred specification includes heteroskedasticity-robust standard errors. To investigate the robustness of our findings with respect to our approach to statistical inference, we use the following alternative approaches reported in Appendix Table A.5: school-level clustering only, school-gender-risk factor clustering, as these were the strata along which treatment status was defined, bootstrapped unclustered standard errors, bootstrapped standard errors with school-level clustering, and bootstrapped standard errors clustered at the school-gender-risk factor level. As shown in Appendix Table A.5, the standard errors for our main point estimates are very similar across these different approaches.

5.3 Heterogeneity by District, Cohort, and Student Characteristics

The broader literature on evaluating programs has increasingly emphasized the salience of treatment heterogeneity. For example, a recent essay by Smith (2022) goes so far as to argue "for the general empirical irrelevance of the common effect model." For this reason, our analysis considers the estimated impact of the CHSS program for subgroups defined by salient and observed student traits in addition to overall impacts averaged across all students. Similarly, we also consider whether there are distinctive program impacts across sites and cohorts (i.e., one District A cohort and two District B cohorts).

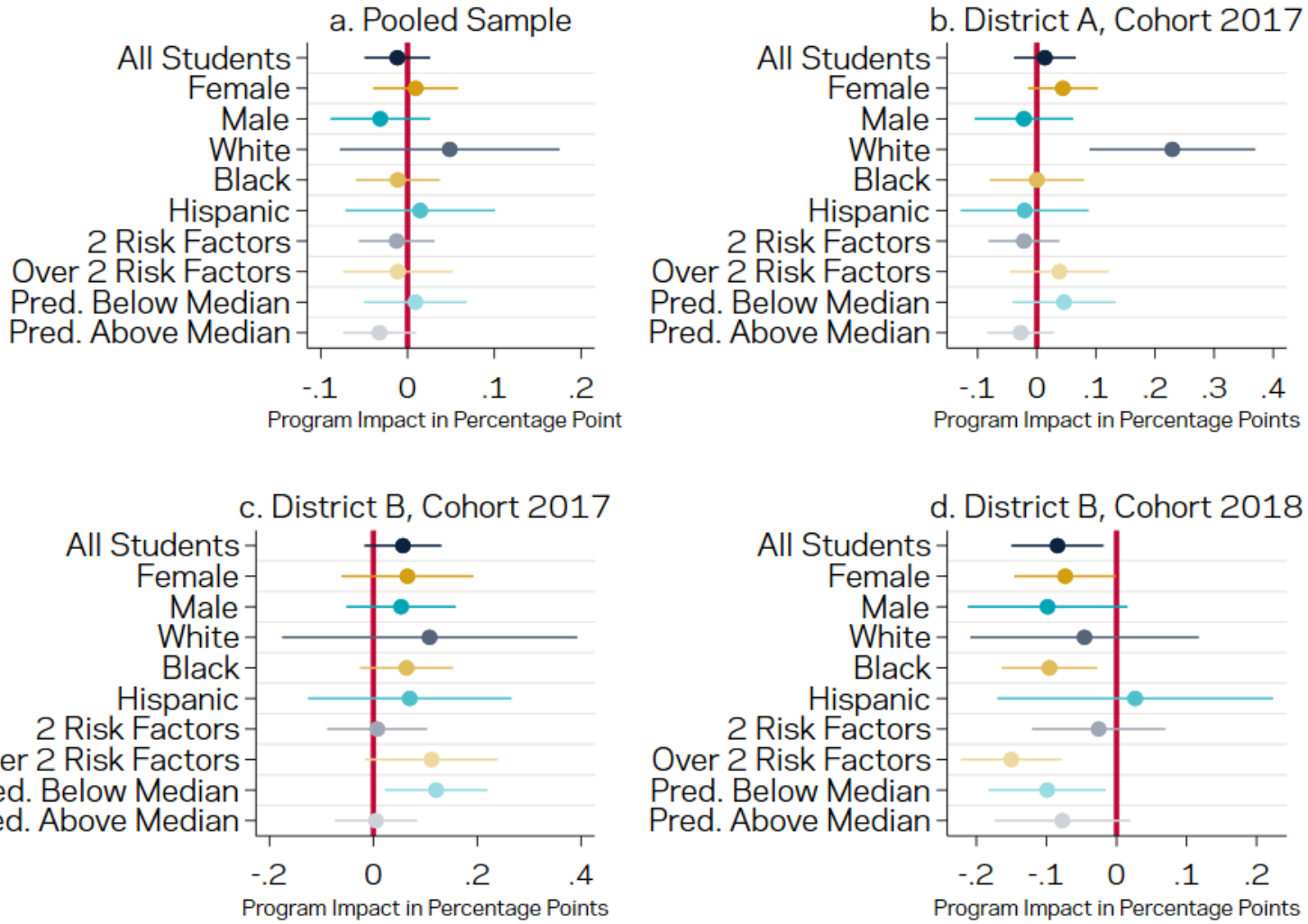
To determine whether the program impacted graduation for certain groups, we estimated treatment effects separately for various demographic groups defined by gender, race and ethnicity, the number of risk factors, and predicted graduation rate.[9] Figure 2 panel (a.) reports the ITT estimates of program impact for several groups for the pooled sample. The dot represents the regression-adjusted difference in graduation rates between the treatment and control group. The horizontal lines through each dot represent the 95 percent confidence interval, calculated using standard errors clustered by school, risk factor, gender, and cohort.

[9]Specifically, for the control group, we regress actual four-year graduation status on the full set of observable characteristics as well as school-level fixed effects, clustering standard errors at the school level.

At the top of the figure, we include the estimate for all students, which is the same as the estimate reported in column (1) of Table 2. These results for the pooled sample show little evidence of an effect of the program for any of the demographic groups–none of the estimates are statistically different from zero. For white students, the point estimate suggests that the program increased graduation rates, but this estimate is very imprecise.

We observe differences in treatment effects across demographic groups within the district and cohort subsamples. Panels (b.) through (d.) of Figure 2 report the results separately by district and cohort. The results for District A (panel (b.)) mirror those for the pooled sample – little evidence of an effect for nearly all demographic groups, but (imprecise) evidence of a positive effect for white students. For the District B 2017 cohort in panel (c.), the treatment effect is positive for all groups. While most of these estimates are imprecise, the treatment effect is large and statistically significant for the most at-risk students–those with the lowest predicted likelihood of graduating high school based on their baseline characteristics. For this group, those assigned to treatment are 4.6 percentage points more likely to graduate within four years than those assigned to the control group. For the 2018 District B Cohort in panel (d.), the point estimate is negative for nearly all subgroups, and for more than half of them, this effect is statistically significant. The negative effect is particularly evident for black students and students with more than two risk factors.

Figure 2: Intent to Treat Estimates of Effect of Program on High School Graduation for Subgroups



Notes: Effects of the program on 4-year high school graduation by subgroups are shown for the pooled sample. This figure uses ever offer as the treatment definition. All estimates include school-level and cohort fixed effects. Standard errors clustered by school/risk factor/gender/cohort. "Pred." represents "Predicted". The predicted below median bars represent the impact of the program for those who were predicted to be below the median graduation rate. The predicted above median bars represent the impact of the program for those who were predicted to be above the median graduation rate. The bars represent 95% confidence intervals.

Appendix Table A.3 provides ITT and treatment-on-the-treated (TOT) estimates for the effect of the program on four-year graduation for the pooled sample, and broken down by district and cohort. The ITT estimates are the same as those reported for "All Students" in Figure 2 panel (a.). The TOT estimates effectively re-scale the ITT estimates to account for the fact that not all students (only about 35 percent) who are assigned to the treatment group take up the program. The TOT for the 2017 District B Cohort indicates that the program increased graduation rates for those who participated in the program by 15.8 percentage points (21 percent). Appendix Table A.3 provides ITT and TOT estimates for the effect of the program on five-year graduation, however the effect of the program does not differ substantively from the effect on four-year graduation rates.

5.4 Understanding District and Cohort Heterogeneity

While we find null results for our full, pooled sample, we find evidence suggesting that the program had very different effects across districts and cohorts. To explore further why the treatment effect might differ by district and cohort, we consider other factors that differed along these dimensions. Specifically, we examine whether differences in how the program was implemented or the effect the program had on intermediate outcomes might shed light on differences in treatment effects.

A conceptual framework introduced by Weiss et al. (2014) states the possible variation in program impact can be organized around the "Three C's": client characteristics, program context, and the treatment contrast. With regard to site-cohort treatment heterogeneity, we expect few differences that might be related to client characteristics or program contexts. The structured eligibility for the CHSS program ensures the similarity of the participating students. Similarly, because our two participating sites are both geographically adjacent and broadly similar in terms of socioeconomic and demographic traits, there is little reason to expect meaningful differences in program contexts.

However, both the multi-faceted nature of the CHSS program, variation in the takeup of offered services, and data on the varied character of the CHSS implementation strongly suggest that participating students experienced starkly different treatment contrasts across sites and cohorts.

5.4.1 Enrollment Timing

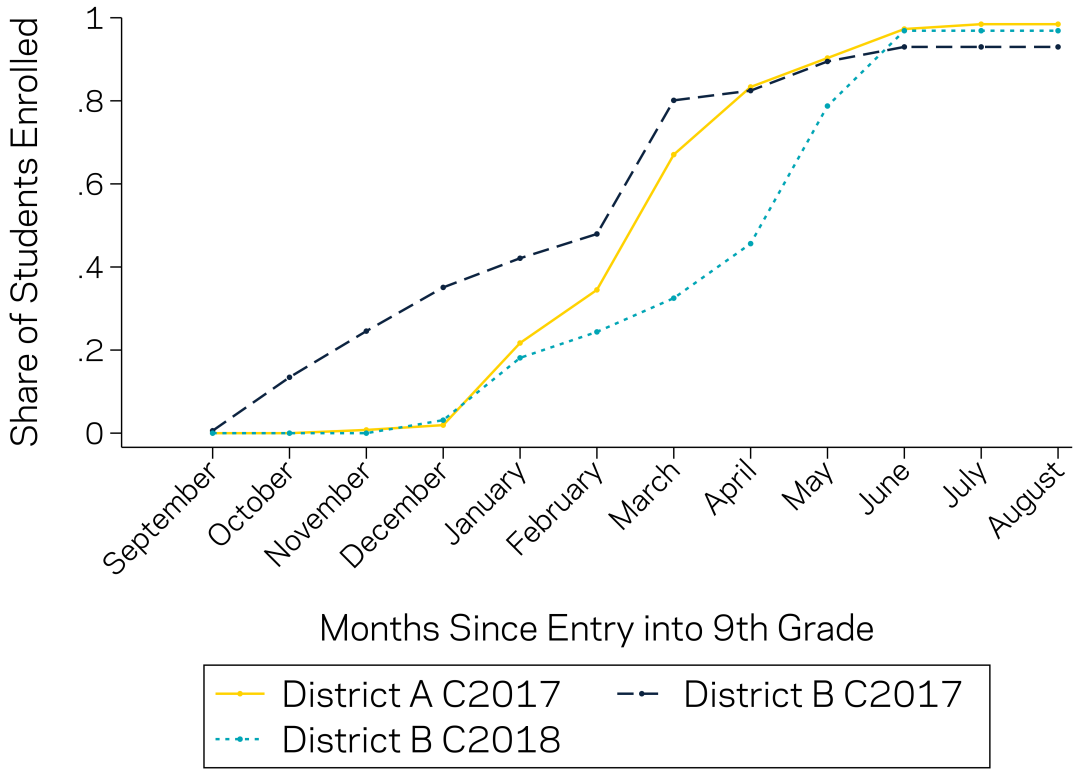
One of the more noticeable differences in treatment contrast across districts and cohorts was the timing of enrollment. The goal of the CHSS program was to enroll at-risk students into the program as early on in their high school experience as possible in order to address issues that might prevent students from being on track to graduate before students get too far off track. That is why eligibility for the program was based on factors prior to enrollment into high school. In fact, the program worked with many of the grade schools in each of the districts to identify and enroll the most at-risk students into the program as they entering high school.

As is shown in Figure 3, there is considerable variation in the timing of enrollment across districts and cohorts, with the 2017 cohort in District B enrolling students much earlier. This figure reports the share of all students who eventually enroll into the CHSS program in a particular district and cohort who have enrolled by each month since starting ninth grade. For the 2017 Dictrict B cohort, over a third of the students who would enroll in CHSS had already enrolled by December of freshman year, which is in sharp contrast to District A or the 2018 cohort District B, neither of which had enrolled even five percent of eventual enrollees by December. The slow enrollment for the 2018 cohort of District B is particularly noticeable as less than half of the eventually enrolled students had enrolled by April of their freshman year.

There were a number of reasons for the unusually slow enrollment of students into CHSS for the 2018 cohort of District B. For starters, data transfer issues delayed the creation of the lists of eligible students, so the case managers which in turn delayed outreach by case managers. In addition, the program faced significant staffing and implementation challenges in this cohort. In an effort to ensure students were enrolled during their ninth-grade year, some students were enrolled before a designated case manager was assigned. As a result, many received only limited services early on and experienced delays in accessing core components of the program such as mentoring and academic support.

Regardless of the reason for the delay, these data suggest that enrollment timing could be an important factor for determining program impact–we see see suggestive program impacts for the one cohort-district that enrolled students early on in their first year, but little evidence of impact, or even a negative impact, in district-cohorts where students were enrolled much later on.

Figure 3: Share of students who ever enrolled in program who have entered, by month since entry into 9th grade and by district.



Note: The share of students who have ever enrolled into the program are shown, by month since entry into 9th grade and distinguished by district. Program entry is defined by program start date provided by program staff.

5.4.2 Other Variation in Program Implementation

Using additional program data from the social service organization that runs CHSS, we explored other differences in the way in which the program was implemented across districts and cohorts. Even though the program has very detailed procedures and manuals, and offers a consistent training program for their case managers, inevitably differences in implementation arise. Appendix Table A.6 reports the fraction of students who enrolled in CHSS that participated in various program activities by district and cohort. We also examined differences in treatment contrast by looking at service engagement on the intensive margin–how many minutes did the students spend engaged in each of these activities (Table 3).

The most obvious difference across these groups is in tutoring. 85 percent of students

enrolled in the 2017 cohort in District B participated in tutoring, which is 18 percentage points more than the participation rate for enrolled students in District A and 25 percentage points more than the rate for those in the 2018 cohort of District B. Also, while nearly all students participated in at least some mentoring services, engagement was much greater for the students in the 2017 cohort in District B than for the other students (Panel A of Table 3). Specifically, the 2017 District B students spend 230 more minutes (15 percent) in mentoring than students in District A, and 412 more minutes (24 percent) than those in the 2018 cohort of District B. Similarly, while the program had contacted the family for nearly all students in the program (Appendix Table A.6), the time spent connecting with family was noticeably greater in the 2017 District B cohort than the other district/cohort (Table 3). For all three of these activities–tutoring, mentoring, and family contact–the district-cohort that engaged students the most in these services was also the district with the largest treatment effect.

There were also some activities that were less emphasized in the 2017 District B cohort. For example, students in this group spent much less time on after-school activities, the category that accounted for the most time for all three groups, than students in District A.

5.4.3 Impact on Early Outcomes

These and other treatment contrast differences across districts and cohorts may have impacted student trajectories well before the four-year graduation mark. For example, if enrolling students early on in high school is important for the program to have an impact on graduation, then you might expect the program to have an impact on early indicators of being on track. So we also examined how the program impacted early and intermediate outcomes, and whether there was any evidence that program impacts for these outcomes differed across cohorts and districts.

To earn a high school diploma, students must meet credit requirements, with a total of 22 credits required to graduate. So, falling behind on credits early on in high school could make it challenging to graduate on time, or at all. Table 4 panel (a) reports estimates of the impact of CHSS on credit accumulation in Year 1. While the treatment effect for the pooled sample is small and not statistically different from zero, we see a positive (8 percent) and marginally significant effect on credits accumulated for the 2017 District B cohort, but a negative and imprecise effect for the 2018 District B cohort. Thus, some of the heterogeneity in treatment effects across districts and cohorts that we see for graduation, is already evident

Table 3: Minutes for Each Activity, Program Enrollees

	Dist. A C17	Dist. B C17	Dist. B C18	Dist. A C17 - Dist. B C17	Dist. B C18 - Dist. B C17
	Means			Means (Standard Errors)	
	(a) Minutes, Unconditional on Participation in Activity				
Tutoring Sessions	505.9	1250.9	305.9	-745.0***	-945.0***
				(146.4)	(139.9)
After School Activities	2617.2	1954.3	1794.8	663.0*	-159.4
				(371.8)	(329.5)
Post HS Exploration	109.0	22.43	25.79	86.6***	3.37
				(24.5)	(10.7)
Mentoring Sessions	1496.6	1726.7	1314.5	-230.1**	-412.2***
				(107.8)	(128.5)
After School Activities	202.1	40.59	44.99	161.5***	4.40
				(28.7)	(14.3)
Employment Related Activities	31.42	53.58	29.05	-22.2	-24.5
				(21.7)	(26.2)
Contacting Family	110.5	318.3	180.0	-207.8***	-138.3***
				(21.2)	(23.2)
Student Enrichment	532.9	65.89	94.94	467.0***	29.0
				(76.6)	(42.3)
Other/Unknown	1841.3	1787.0	1624.3	54.3	-162.7
				(340.8)	(313.5)
Total Obs.	257	169	158		
	(b) Minutes, Conditional on Participation in Activity				
Tutoring Sessions	755.8	1468.1	508.7	-712.2***	-959.3***
				(171.1)	(161.0)
After School Activities	2924.5	2752.2	2675.3	172.2	-76.9
				(431.2)	(435.2)
Post HS Exploration	329.6	126.3	239.7	203.3***	113.4
				(70.6)	(69.6)
Mentoring Sessions	1496.6	1747.4	1331.3	-250.8**	-416.0***
				(107.7)	(128.7)
After School Activities	509.2	228.7	222.2	280.5***	-6.51
				(64.0)	(53.6)
Employment Related Activities	118.8	905.5	573.8	-786.8***	-331.7
				(226.8)	(340.7)
Contacting Family	117.3	324.0	182.3	-206.7***	-141.7***
				(21.3)	(23.3)
Student Enrichment	925.3	232.0	441.2	693.4***	209.2
				(139.1)	(170.7)
Other/Unknown	2170.7	2626.1	2491.7	-455.4	-134.4
				(415.8)	(427.7)
Total Obs.	257	169	158		

Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. Sample includes students that enrolled into the program. Differences are not regression-adjusted.

by the end of the first year of high school. We also looked at credit accumulation beyond the first year, but due to data limitations we only observe these outcomes for District A.

Table 4: Effect of Program on Credit Accumulation, Pooled and by District - Analytic Sample

	Pooled	District A	District B	
	ITT	ITT	ITTC17	ITTC18
	(a) Credits in Year 1			
	(1)	(2)	(3)	(4)
Treatment	-0.008	0.034	0.373*	-0.288
	(0.114)	(0.169)	(0.218)	(0.209)
	[0.942]	[0.842]	[0.089]	[0.169]
Control Mean	4.888	5.432	4.38	4.714
N	2339	932	676	731
	(b) Credits Accumulated in Years 1 and 2			
	(1)	(2)	(3)	(4)
Treatment	0.593	0.593	0.000	0.000
	(0.402)	(0.402)	(.)	(.)
	[0.141]	[0.141]	[.]	[.]
Control Mean	16.87	16.87	0	0
N	834	834	693	736
	(c) Above Credit Threshold			
	(1)	(2)	(3)	(4)
Treatment	0.006	0.006	0.030	-0.044
	(0.029)	(0.029)	(0.039)	(0.035)
	[0.824]	[0.824]	[0.436]	[0.214]
Control Mean	.6132	.6132	.5088	.5441
N	932	932	676	731

Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Ever offer is used as the treatment definition. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. All regressions include school-level and cohort fixed effects. The sample is a subset of the 2,381 students in the analytic sample with observable graduation outcomes that also had observable credit accumulation. Credit accumulation was only available in District A for panel (b) due to data issues in District B. Outcomes are conditional on enrollment in a school year for at least 10 days and attendance for at least one of those days.

Another important early indicator of success in high school is absenteeism. Absenteeism is associated with negative outcomes including low academic achievement, school truancy,

dropping out, delinquency, and substance abuse. Moreover, the negative effects of chronic absenteeism can spill over to regularly attending students as teachers divert their attention to meet the learning and social needs of chronically absent children when they return to school. We use the State Education Department's definition of chronic absenteeism, which is defined as missing 10 percent or more of school over an academic year for any reason (excused or unexcused). We examine this outcome for each year that a student is enrolled, restricting the sample to those who are enrolled at a school for at least ten days and present for one of them.

Table 5 reports intent to treat estimates of the program impact on chronic absenteeism for the pooled sample as well as separately by district and cohort. For the pooled sample the intervention led to a 2.5 percentage point (7 percent) decline in chronic absenteeism in the first year; however, this effect is not significant. The treatment effect is most noticeable for the 2017 cohort in District B, where the intervention led to a 5.8 percentage point (16 percent) decline in absenteeism. These effects seem to attenuate a bit through year 2 and are less precisely estimated.[10] We find little evidence that the program led to reduced absenteeism for the 2018 District B cohort.

[10] Attendance data are missing for students in the District B school district for the 2020-2021 academic year, year 4 for the 2017 cohort, and year 3 for the 2018 cohort.

Table 5: Effect of Program on Chronic Absenteeism, Pooled and by District - Analytic Sample

	Pooled	District A	District B	
	ITT	ITT	ITTC17	ITTC18
	(a) Chronically Absent in Year 1			
	(1)	(2)	(3)	(4)
Treatment	-0.025	-0.031	-0.058*	0.010
	(0.019)	(0.030)	(0.035)	(0.032)
	[0.178]	[0.301]	[0.098]	[0.748]
Control Mean	.3605	.4625	.3593	.2628
N	2349	936	679	734
	(b) Any Chronic Absenteeism in Years 1 or 2			
	(1)	(2)	(3)	(4)
Treatment	-0.009	-0.033	-0.026	0.025
	(0.019)	(0.029)	(0.037)	(0.035)
	[0.653]	[0.252]	[0.482]	[0.474]
Control Mean	.5056	.6161	.4549	.4337
N	2362	942	684	736

Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Ever offer is used as the treatment definition. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. All regressions include school-level and cohort fixed effects. The sample is a subset of the 2,381 students in the analytic sample with observable graduation outcomes that also had observable chronic absenteeism. Chronic absenteeism data is unavailable in District B Cohort 17 for year 3 and District B Cohort 18 for year 4 due to COVID-19. Both outcomes are conditional on enrollment in a school year for at least 10 days and attendance for at least one of those days.

Another important indicator for being on track, particularly for the state where this intervention was implemented, is whether the student sat for and/or passed the state's required standardized exams for core courses. Passage of these exams was required for a standard

high school diploma in this state and students who did not pass these exams early on in high school are much less likely to graduate. In Table 6, we report the treatment effects on whether the student took at least one of the exams and whether they passed at least one. While the point estimates are positive for the pooled sample (column 1), these estimates are small and imprecisely estimated. For the 2017 District B cohort, however, we see a sizable impact of the program–those in the treatment group were 9.8 percentage points (13 percent) more likely to take a core exam and 6.6 percentage points more likely to pass one than students in the control group; the former estimate is statistically significant at the 99 percent level, while the latter estimate is marginally significant. For District A, the treatment effect is positive but imprecisely measured, and for the 2018 District B cohort the effect is negative but imprecisely measured.

Table 6: Effect of Program on Taking and Passing Standardized Exams in First Year, Pooled and by District - Analytic Sample

	Pooled	District A	District B	
	ITT	ITT	ITTC17	ITTC18
	(a) Took Standardized Exam			
	(1)	(2)	(3)	(4)
Treatment	0.023	0.044	0.098***	-0.047
	(0.018)	(0.030)	(0.032)	(0.031)
	[0.196]	[0.140]	[0.002]	[0.128]
Control Mean	.6325	.3921	.7468	.7892
N	2380	951	693	736
	(b) Passed at Least 1 Standardized Exam			
	(1)	(2)	(3)	(4)
Treatment	0.017	0.036	0.066*	-0.043
	(0.019)	(0.029)	(0.036)	(0.033)
	[0.357]	[0.210]	[0.073]	[0.195]
Control Mean	.3875	.3343	.3755	.4488
N	2380	951	693	736

Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Ever offer is used as the treatment definition. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. The sample is a subset of the 2,381 students in the analytic sample with observable graduation outcomes that also had observable standardized exam data. All regressions include school-level and cohort fixed effects.

In addition to examning specific early indicators, we assess treatment effects on two broader meaures of student progress: persistence in high school and a composite on-track index. Consistent with our pre-analysis plan, we examine persistence through binary indicators of whether a student dropped out by the end of their first, second, or third year of high school. While dropout is rare in the first year due to compulsory schooling laws, rates begin to

diverge in later years.

Taking the evidence of the impact of CHSS on early outcomes together, we see that across all of these predictors of being on track to graduate, the district and cohort where we find suggestive evidence of a positive effect on graduation, is also the group for which we see positive impacts on early outcomes. This suggests that working with students and addressing barriers early on in high school might be an important mechanism by which this program improves high school completion.

5.5 Post-Secondary Outcomes

While a primary goal of the CHSS program is to promote college completion, core components of the program focus on preparing students for success both in the labor market and in post-secondary schooling. We are not able to assess the impact of the program on post-high school labor market outcomes due to data restrictions, but we do have access to post-secondary data that allow us to examine the impact of the program on educational outcomes beyond high school.

We obtained college enrollment data from the National Student Clearinghouse for all students in our study who were uniquely identifiable by first name, last name, and date of birth, which was nearly the entire sample. [11] The NSC data cover post-secondary enrollment through December 2025, which is seven years after our 2018 cohort was enrolled. From these NSC data, we constructed several measures of post-secondary education including: whether a student ever enrolls in college; whether a student enrolls within one-year of high school; the total number of semesters enrolled conditional on being enrolled; and post-secondary degree completion.

We note that the expected effect of CHSS on post-secondary education is ambiguous. On the one hand, the program worked with students to be prepared for schooling beyond high school, including help with SAT/ACT registration, college applications, and scholarship acquisition, as well as organized college visits. This might lead to increased post-secondary enrollment. On the other hand, the program also prepared students for success in the labor market after high school by offering resume support, interview preparation, and job readi-

[11] The National Student Clearinghouse is a nonprofit organization founded in 1993 that provides nationwide verification of student enrollment and degree attainment. As of today, the NSC works with institutions enrolling approximately 97% of U.S. college students.

ness workshops, along with opportunities for job shadowing, career exploration and speaker events. These components may have encouraged some students to pursue employment, potentially reducing post-secondary enrollment.

As shown in Table 7, we do not find evidence that the program had an impact on students' likelihood of enrolling in college for the full sample. Nor do we find evidence of an effect on college enrollment for any of the subgroups we considered (Table 7). We also looked at whether a student is enrolled in college within one year of high school graduation, because if the program were to affect enrollment it would arguable have its most noticeable effect immediately following exiting the program. Here, again, we find little evidence that the program increased post-secondary enrollment. In fact, there is some evidence that the program reduced enrollment initially. But because this effect is driven by the 2018 District B cohort, this may be a result of the fact that, as shown above, for this district-cohort those in the treatment group were less likely to graduate high school.

We find some evidence that the program increases the total number of semesters that a student enrolls at a post-secondary institution, conditional on having enrolled for at least one semester (Table 7, panel b). Students in the treatment group were on average enrolled in an additional 0.4 semesters compared to the control group, and this effect is marginally significant. The effect is most noticeable for the District A cohort, where students who were offered the treatment enrolled in an additional 0.94 semesters.

Finally, we explored the effect of the program on post-secondary degree completion. We should caution that because we only have data for either seven or eight years after initial enrollment in the program, only students who enrolled soon after graduating high school and who maintained a fairly full course load would have the opportunity to have completed a two-year post-secondary degree within this period. This is evident in the means for the control group — less than 2.5 percent of the control group in the pooled sample has completed a degree within this time frame. Our estimates suggest that the program had little effect on degree completion for the pooled sample. There is some evidence that the program increased post-secondary graduation in District A. Those in the treatment group were 1.8 percentage points more likely to have graduated, and this effect is marginally significant.

Table 7: Effect of Program on Post-Secondary Outcomes, Pooled and by District - Analytic Sample

	Pooled	District A	District B	
	ITT	ITT	ITTC17	ITTC18
	(a) Ever Enrolled in Post-Secondary Education as of December 2024			
	(1)	(2)	(3)	(4)
Treatment	-0.067***	-0.049*	-0.036	-0.112***
	(0.019)	(0.030)	(0.038)	(0.034)
	[0.000]	[0.098]	[0.345]	[0.001]
Control Mean	.3899	.3344	.4093	.4307
N	2370	941	693	736
	(b) Enrolled in Post-Secondary Education Within 1 Year of Graduation			
	(1)	(2)	(3)	(4)
Treatment	-0.032*	0.001	-0.023	-0.079**
	(0.018)	(0.027)	(0.036)	(0.033)
	[0.079]	[0.958]	[0.522]	[0.019]
Control Mean	.3184	.2147	.3713	.3825
N	2370	941	693	736
	(c) Total Semesters Ever Enrolled as of December 2024			
	(1)	(2)	(3)	(4)
Treatment	0.381*	0.937**	0.219	-0.242
	(0.210)	(0.365)	(0.436)	(0.306)
	[0.070]	[0.011]	[0.617]	[0.430]
Control Mean	3.931	3.257	4.68	3.937
N	819	285	264	270
	(d) Completed Post-Secondary Education as of December 2024			
	(1)	(2)	(3)	(4)
Treatment	0.011	0.018*	0.015	0.000
	(0.007)	(0.009)	(0.016)	(0.013)
	[0.108]	[0.055]	[0.338]	[0.976]
Control Mean	.02458	.01227	.03376	.03012
N	2370	941	693	736

Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. All regressions include school-level and cohort fixed effects. The sample is all of the students in our analytic sample that also were able to match to National Student Clearinghouse data due to non-missing first names, last names, and dates of birth which are needed to match. NSC data is available through December 31, 2024. On-time graduation for students in the 2017 cohort is defined as spring, 2021. On-time graduation for students in the 2018 cohort is defined as spring, 2022. Ever offer is used as the treatment definition. Total semesters enrolled is conditional on having enrolled into post-secondary education.

VI Discussion

On average, the CHSS program had little or no detectable effect on four- or five-year high school graduation rates. Both the overall intent-to-treat estimates and treatment-on-the-treated estimates are small and statistically indistinguishable from zero, suggesting that, on average, the program did not significantly alter graduation outcomes. At first glance, this null finding could be interpreted as evidence of program ineffectiveness. However, in this section, we assert that the aggregated result masks substantial variation across cohorts, districts, and student subgroups that highlight the importance of considering mechanisms and context through which the program operates, and serves as an important warning that, even in an RCT context where classical theory dictates that the difference in observed outcome can be attributed entirely to the program itself, that consideration of mechanisms and the context through which the program operates can be equally as important.

Examining results by site and cohort highlights these differences. The District B 2017 cohort experienced large and positive effects on high school graduation, with treated students more likely to complete high school on time. Conversely, the District B 2018 cohort exhibited negative and statistically significant effects, with treated students less likely to graduate than their counterparts. District A, in contrast, shows no discernible effect. These patterns underscore that the effectiveness of a comprehensive wraparound intervention is highly contingent on the local implementation environment and timing, emphasizing the centrality of the program context and fidelity.

Heterogeneity by student characteristics further illustrates the nuanced nature of program impacts. In some contexts, the most at-risk students, those with the lowest predicted likelihood of graduating, appear to benefit meaningfully from the intervention, particularly when implementation is timely and intensive.

Patterns in credit accumulation, chronic absenteeism, and exam-taking mirror the graduation results: the cohort that enrolled early and received robust services (District B 2017) shows gains across these intermediate outcomes, whereas delayed or fragmented enrollment (District B 2018) coincides with stagnant or negative outcomes. Table 4 shows that students in the District B 2017 cohort attained an additional 0.37 credits in the first year, an 8.5 percentage point increase compared to the control group mean. In contrast, the District B 2018 cohort attained 0.288 credits less in the first year as a result of the program, a 6.1 percentage-point reduction from the control group mean. These patterns continue through

other observable intermediary outcomes, such as chronic absenteeism, and taking and passing at least one standardized exam, with the District A 2017 cohort not experiencing any statistically significant differences as a result of the program. These early measures suggest that program impacts on graduation operate through the accumulation of timely, sustained support, reinforcing the notion that service intensity and engagement are crucial drivers of success.

The absence of a general treatment effect conceals sharp differences in outcomes between cohort years. Although pooled estimates are indistinguishable from zero, disaggregated analyses show that both the year and month of enrollment are critical determinants of program impact: some cohorts experienced substantial positive effects, while others saw negative outcomes.

In particular, the District B 2017 cohort, where nearly 40 percent of students were enrolled in the program by December, compared with almost none in the District A 2017 and District B 2018 cohorts, performed significantly better on both intermediate and terminal outcomes. This pattern is consistent with evidence that many high school friendships form within the first few months and are strongly shaped by structural proximity (Hays, 1985; Frank et al., 2014). Early enrollment may therefore have facilitated strong early-term friendship development, strengthening both peer networks and program engagement. In contrast, the District B 2018 cohort faced delays in enrollment, staff turnover, and inconsistent delivery of key services, coinciding with negative effects. This temporal heterogeneity highlights how differences in timing and context interact with client characteristics to shape outcomes, in line with the Three C's framework.

Beyond within-semester variation, cohort-level timing is also consequential. The 2018 cohort encountered the onset of the COVID-19 pandemic earlier in their high school trajectory, suggesting that, even under earlier program enrollment, service delivery may still have been substantially disrupted. Distinguishing between effects driven by month-level enrollment timing and those attributable to year-specific shocks is inherently difficult, as the data only include a single district-year cohort for 2018.

These findings also reinforce the theory that comprehensive interventions frequently face an impact-scale trade-off: as programs broaden their reach or confront operational challenges, maintaining high levels of implementation fidelity, and therefore effectiveness, becomes increasingly difficult (Mcdonald et al., 2006; Kincald and Horner, 2018). Even modest deviations in staff capacity, service intensity, or coordination can markedly diminish the effec-

tiveness of wraparound supports, particularly for vulnerable students who benefit most from early and continuous engagement. In District A, CHSS faced high staff turnover and related capacity challenges in 2017, which constrained the consistency of service delivery. Structural differences between districts further shaped implementation. District B maintained a designated program space that facilitated in-person tutoring by local college students, provided snacks, and fostered peer interaction among participants. By contrast, District A lacked a dedicated space, and tutoring was more often conducted remotely, limiting both intensity and opportunities for social connection. The heterogeneity observed across cohorts underscores that program scale and implementation constraints interact with client characteristics in shaping realized outcomes.

Treatment contrast — the difference between what treatment and control students actually experience — emerges as a key driver of results. The benefits of CHSS are evident only when treated students receive services that exceed those available to control students. During the COVID-era, for instance, many control students may have accessed alternative supports such as tutoring, mentoring, or online academic enrichment, narrowing the effective treatment contrast. This convergence may partially explain muted overall impacts and highlights how evolving external conditions can modulate program effectiveness, underscoring the importance of continuously monitoring the gap between intended and realized services.

Evaluations from a single site or cohort may be misleading if used to generalize conclusions. The sharp differences observed across cohorts and districts illustrate that context matters: evidence should inform refinements in implementation and operational design rather than be taken as a simple verdict on program efficacy. Scaling programs requires careful attention to local capacity, operational constraints, and student needs to ensure that benefits observed in early, well-resourced implementations are maintained at scale.

Moreover, the heterogeneity in program effects underscores the challenge of scaling comprehensive wraparound interventions. Programs that rely on timely, intensive, and individualized supports may not deliver consistent benefits without strong systems for implementation monitoring, staff training, and continuous quality improvement. Policymakers should therefore exercise caution in interpreting null or negative results from large-scale deployments, recognizing that apparent ineffectiveness may reflect implementation failures rather than a fundamental flaw in program design.

This study adds to the growing evidence on the inverse relationship between program scale and impact, illustrating that fidelity-driven explanations, rather than selection bias,

can account for heterogeneous outcomes across cohorts (Angrist and Meager, 2023). By documenting sharp differences in program effectiveness across districts and cohorts, we show that even well-designed interventions may fail to produce consistent outcomes when implementation fidelity varies.

Second, our findings serve as a cautionary tale about generalizing from small-scale or single-site randomized controlled trials. Programs that show large, positive effects in one context may not scale effectively if local implementation capacity and contextual factors differ, underscoring the limits of external validity for replication studies.

Finally, this study highlights the importance of heterogeneity analyses and continuous improvement frameworks in program evaluation. By connecting observed outcomes to the "Three C's" framework: client characteristics, program context, and treatment contrast, we provide a structured lens for understanding why programs succeed in some contexts but fail in others, offering guidance for both evaluators and practitioners.

VII Conclusion

This study demonstrates that average treatment effects of wraparound supports are fragile. While comprehensive models have the potential to improve graduation outcomes, their success depends critically on local implementation capacity and fidelity. Even well-designed programs may fail to generate consistent effects in the absence of timely recruitment, staff continuity, and service intensity. In this sense, the benefits of wraparound interventions are contingent on fidelity to a flagship model, with realized effects hinging on whether schools can sustain high-quality, early, and coordinated service delivery.

The results also highlight the limited generalizability of much of the existing evidence on programs to reduce high school dropout. Single-site or single-cohort randomized controlled trials may offer compelling causal estimates in context, but these can obscure meaningful variation across settings and years. In the case of CHSS, the same program produced large positive effects in one cohort and negative effects in another *within the same district*. Such divergent findings underscore the need for replication across time and place before claims of effectiveness are made.

References

Alberto Abadie, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. When should you adjust standard errors for clustering? *Quarterly Journal of Economics*, 138:1–35, 2023.

Elaine M. Allensworth. Graduation and dropout trends in chicago: A look at cohorts of students from 1991 through 2004. Technical report, Consortium on Chicago School Research at the University of Chicago, 2005.

Elaine M. Allensworth and John Q. Easton. The on-track indicator as a predictor of high school graduation. Technical report, Consortium on Chicago School Research, 2005.

Joshua D. Angrist and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. *Princeton University Press*, 2009.

Noam Angrist and Rachael Meager. Implementation matters: Generalizing treatment effects in education. *Available at SSRN 4487496.*, 2023.

David Card. The causal effect of education on earnings. *Handbook of Labor Economics*, 1999.

Julie Berry Cullen, Steven D. Levitt, Erin Robertson, and Sally Sadoff. What can be done to improve struggling high schools? *Journal of Economic Perspective*, 2013.

Clément de Chaisemartin and Luc Behaghel. Estimating the effect of treatments allocated by randomized waiting lists. *Econometrica*, 2020.

Thomas S. Dee. The promises and the reality of school reform: Three forces that matter. Policy brief, Hoover Institution, 2024.

Kenneth A Frank, Chandra Muller, and Anna S Mueller. The embeddedness of adolescent friendship nominations: The formation of social capital in emergent network structures. *American Journal of Sociology*, 2014.

Jonathan Guryan, Sandra Christenson, Ashley Cureton, Ijun Lai, Jens Ludwig, Catherine Schwarz, Emma Shirey, and Mary Clair Turner. The effect of mentoring on school attendance and academic outcomes: A randomized evaluation of the check connect program. *Journal of Policy Analysis and Management*, 2020.

R B Hays. A longitudinal study of friendship development. *Journal of personality and social psychology*, 1985.

Micha D. Segeritz Kemple, James J. and Nickisha Stephenson. Building on-track indicators for high school graduation and college readiness: Evidence from new york city. *Journal of Education for Students Placed at Risk*, 18:7–28, 2013.

Don Kincald and Rob Horner. Changing systems to scale up an evidence-based educational intervention. *Evidence-Based Communication Assessment and Intervention*, 2018.

Sarah-Kathryn Mcdonald, Vanessa Ann Keesler, and Barbara Schneider. Scaling-up exemplary interventions. *Educational Researcher*, 2006.

National Center for Education Statistics. Digest of education statistics, table 219.10: High school graduates, by sex and control of school; public high school averaged freshman graduation rate (afgr); public high school 4-year adjusted cohort graduation rate (acgr); and total graduates as a ratio of 17-year-old population: Selected school years, 1869-70 through 2022-23. Online, 2025. URL `https://nces.ed.gov/programs/digest/d24/tables/dt24_219.10.asp?current=yes`. Accessed July 14, 2025.

Jeffrey Smith. Treatment effect heterogeneity. *Evaluation Review*, 2022.

Michael J. Weiss, Howard S. Bloom, and Thomas Brock. A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, 2014.

A Appendix

Table A.1: Regression Adjusted Comparison of Missingness, Pooled Sample

	Full Sample	Control	Treatment	Difference
	Means	Means	Means	Treatment - Control
Missing Grad Outcome	0.0581	0.0577	0.0584	0.00029
				(0.0097)
Transferred Outside of NY State	0.0186	0.0241	0.0152	-0.0079
				(0.0059)
Unknown Reason for Missingness	0.0403	0.0336	0.0444	0.0095
				(0.0079)
Match to NSC	0.989	0.993	0.987	-0.0044
				(0.0037)
In Analytic Sample	0.942	0.942	0.942	-0.00029
				(0.0097)
Match to NSC and in Analytic Sample	0.938	0.939	0.937	-0.0027
				(0.0099)
Total Obs.	2,528	953	1,575	2,528

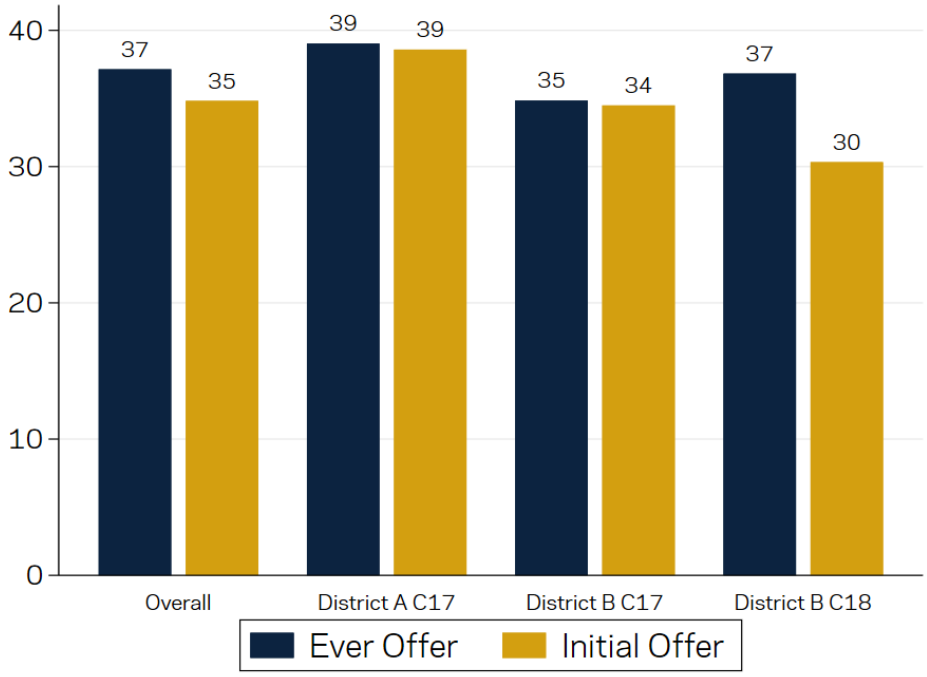
Note: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. Heteroskedasticity robust standard errors are used. Ever offer is used as the treatment definition. Control variables include binary indicators for whether the student is female and whether the student has over two risk factors. Missingness is defined as students with transfers to other schools or high-school equivalency programs, unknowns, or other ambiguous statuses. First names, last names, and dates of birth were required to successfully match to National Student Clearinghouse (NSC) records.

Table A.2: Regression Adjusted Comparison of Missingness, by District and Cohort

	Control	Treatment	Total (District A)	Control	Treatment	Total (District B C17)	Control	Treatment	Total (District B C18)
	Means	Means	Treatment - Control	Means	Means	Treatment - Control	Means	Means	Treatment - Control
Missing Grad Outcome	0.247	0.212	-0.035	0.252	0.206	-0.045	0.189	0.254	0.064**
			(0.028)			(0.033)			(0.030)
Transferred to HSE Program	0.0116	0.0121	0.00015	0.0551	0.0515	-0.0030	0.0451	0.0769	0.032*
			(0.0072)			(0.017)			(0.017)
Transfer	0.122	0.112	-0.010	0.142	0.120	-0.022	0.0423	0.0606	0.019
			(0.022)			(0.026)			(0.016)
Unknown	0.108	0.0832	-0.023	0.0512	0.0330	-0.018	0.101	0.117	0.013
			(0.020)			(0.016)			(0.022)
Total Obs.	344	661	1,005	254	485	739	355	429	784

Note: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. Heteroskedasticity robust standard errors are used. Ever offer is used as the treatment definition. Control variables include binary indicators for whether the student is female and whether the student has over two risk factors. Missingness is defined as students with transfers to other schools or high-school equivalency programs, unknowns, or other ambiguous statuses.

Figure A.1: The fraction of students assigned to treatment who enrolled in the program



Note: Figure reports the share of students that were assigned to the treatment group that took up the program. Two treatment definitions are reported. Enrollment is defined as having a program start date.

Table A.3: Effect of Program on 4 and 5-Year Graduation, Pooled and by District

	Pooled		District A		District B			
	ITT	TOT	ITT	TOT	ITTC17	TOTC17	ITTC18	TOTC18
	(a) 4-Year Graduation							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.012		0.014		0.057*		-0.084***	
	(0.018)		(0.027)		(0.033)		(0.032)	
	[0.508]		[0.619]		[0.087]		[0.008]	
Enrolled into Treatment		-0.031		0.033		0.160*		-0.244***
		(0.047)		(0.066)		(0.093)		(0.095)
		[0.507]		[0.615]		[0.085]		[0.010]
Control Mean	.7294		.7143		.7257		.747	
Control Complier Mean		.891		.8201		.72		1.092
N	2380	2380	951	951	693	693	736	736
	(b) 5-Year Graduation							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.004		0.021		0.060*		-0.071**	
	(0.017)		(0.026)		(0.033)		(0.031)	
	[0.837]		[0.421]		[0.069]		[0.021]	
Enrolled into Treatment		-0.009		0.052		0.169*		-0.206**
		(0.046)		(0.064)		(0.092)		(0.091)
		[0.837]		[0.414]		[0.067]		[0.023]
Control Mean	.7572		.7508		.7342		.7801	
Control Complier Mean		.8849		.8251		.7113		1.074
N	2380	2380	951	951	693	693	736	736

Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. All regressions include school-level and cohort fixed effects. These results are for our analytic sample, which includes the 2,381 students with observable graduation outcomes. Ever offer is used as the treatment definition.

Table A.4: Robustness to Treatment Definition

	Pooled	District A	District B	
	C17 & C18	C17	C17	C18
	(a) 4-Year Graduation			
	(1)	(2)	(3)	(4)
Treatment	**-0.010**	0.015	0.051	-0.082
	(0.018)	(0.027)	(0.033)	(0.033**)
	[0.022]	[0.027]	[0.034]	[0.044*]
Control Mean	**0.7332**	0.7108	0.7308	0.7654
N	**2,381**	952	693	736
	(b) 5-Year Graduation			
	(1)	(2)	(3)	(4)
Treatment	**-0.001**	0.026	0.054	-0.071
	(0.017)	(0.027)	(0.033)	(0.032**)
	[0.021]	[0.023]	[0.034]	[0.045]
Control Mean	**0.7594**	0.7446	0.7393	0.7984
N	**2,381**	952	693	736

Notes: Bolded analyses were pre-registered. *** p<0.01 ** p<0.05 * p<0.1. All estimates include school-level and cohort fixed effects. ( ): robust unclustered standard errors, [ ]: standard errors clustered by school/risk factor/gender/cohort. This table uses initial offer as the treatment definition. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. These results are for our analytic sample, which includes the 2,381 students with observable graduation outcomes.

Table A.5: Robustness to Clustering of Standard Errors

	(1) Pooled	(2) District A C17	(3) District B C17	(4) District B C18
Graduated within 4 Years				
Treatment	-0.012	0.014	0.057	-0.084
	(0.0175)	(0.0273)	(0.0333*)	(0.0317***)
	[0.0197]	[0.0292]	[0.0135**]	[0.0213**]
	((0.0189))	((0.0259))	((0.0354))	((0.0314**))
	[[0.0000]]	[[0.0000]]	[[0.0353]]	[[0.0315**]]
	|0.0169|	|0.0263|	|0.0336*|	|0.0316***|
	<0.0184>	<0.0259>	<0.0136***>	<0.0197***>
	/0.0189/	/0.0250/	/0.0349/	/0.0304***/
Control Mean	0.7294	0.7143	0.7257	0.7470
N	2,381	952	693	736

Notes: *** p<0.01 ** p<0.05 * p<0.1. All estimates include school-level and cohort fixed effects. Coefficients from Table 2 shown in the first row of the table. All other values are standard errors. All Pooled estimates are clustered by cohort except for the unclustered robust standard errors denoted by ( ). Ever offer is used as the treatment definition.

( ): unclustered (robust) standard errors

[ ]: school-level clustering

(( )): school/risk factor/gender clustering

[[ ]]: school/risk factor/gender clustering using cluster-robust variance estimators

| |: bootstrapped unclustered

< >: bootstrapped school-level clustering

/ /: bootstrapped + clustered at the school/risk factor/gender level

Table A.6: Participation Rates for Each Activity, Program Enrollees

	Dist. A C17	Dist. B C17	Dist. B C18	Dist. A C17 - Dist. B C17	Dist. B C18 - Dist. B C17
	Means			Means (Standard Errors)	
Tutoring	0.669	0.852	0.601	-0.18***	-0.25***
				(0.040)	(0.048)
After School Activities	0.903	0.751	0.677	0.15***	-0.074
				(0.038)	(0.050)
Post HS Exploration	0.331	0.178	0.108	0.15***	-0.070*
				(0.042)	(0.038)
Mentorship	1	0.988	0.987	0.012	-0.00082
				(0.0083)	(0.012)
Extracurricular Activities	0.397	0.178	0.203	0.22***	0.025
				(0.042)	(0.044)
Employment Related Activities	0.265	0.0651	0.0506	0.20***	-0.014
				(0.034)	(0.026)
Contacted Family	0.942	0.982	0.987	-0.041**	0.0051
				(0.018)	(0.014)
Student Enrichment	0.580	0.314	0.215	0.27***	-0.098**
				(0.047)	(0.049)
Other/Unknown	0.864	0.722	0.658	0.14***	-0.064
				(0.041)	(0.051)
Total Obs.	257	169	158		

Notes: *** p<0.01 ** p<0.05 * p<0.1. Heteroskedasticity robust standard errors are in parenthesis. Sample includes students that enrolled into the program. Differences are not regression-adjusted.

Table A.7: Baseline Balance Table by District and Cohort, Analytic Sample

	Control	Treatment	Total (District A)	Control	Treatment	Total (District B C17)	Control	Treatment	Total (District B C18)
	Means	Means	Treatment - Control	Means	Means	Treatment - Control	Means	Means	Treatment - Control
White	0.101	0.0674	-0.033*	0.224	0.193	-0.030	0.214	0.228	0.014
			(0.019)			(0.033)			(0.031)
Black	0.530	0.586	0.055	0.553	0.529	-0.025	0.512	0.507	-0.0061
			(0.034)			(0.040)			(0.037)
Hispanic	0.323	0.284	-0.039	0.114	0.171	0.057**	0.130	0.136	0.0069
			(0.032)			(0.027)			(0.025)
English Second Language	0.311	0.300	-0.011	0.215	0.180	-0.035	0.247	0.223	-0.023
			(0.032)			(0.032)			(0.031)
In Poverty	0.948	0.950	0.00040	0.903	0.910	0.0068	0.931	0.943	0.011
			(0.015)			(0.023)			(0.018)
Enrollment Age (Years)	14.13	14.16	0.020	14.08	14.04	-0.038	14.10	14.01	-0.095**
			(0.040)			(0.051)			(0.043)
Predicted Graduation Rate	0.726	0.724	0.0041	0.738	0.736	-0.0010	0.729	0.724	-0.0053
			(0.0047)			(0.0065)			(0.0061)
Female	0.726	0.724	0.0038	0.738	0.736	0.0083	0.729	0.724	0.026
			(0.034)			(0.040)			(0.037)
Over Two Risk Factors	0.726	0.724	0.025	0.738	0.736	0.0048	0.729	0.724	0.0093
			(0.034)			(0.040)			(0.037)
Total Obs.	329	623	952	237	456	693	332	404	736

Note: *** p<0.01, ** p<0.05, * p<0.10. Heteroskedasticity robust standard errors are used. Statistics are taken from the data provided by the school districts. The predicted graduation rate is a predicted value of the graduation rate based on the other covariates in the table. The number of observations for enrollment age and predicted graduation rate is one less for the District B sample due to missing information. Ever offer is used as the treatment definition.

Table A.8: Effect of Program on Being on Track After First Year, Pooled and by District - Analytic Sample

	Pooled	District A	District B	
	ITT	ITT	ITTC17	ITTC18
	(a) Full Sample			
	(1)	(2)	(3)	(4)
Treatment	0.009	0.043*	0.020	-0.040
	(0.017)	(0.025)	(0.034)	(0.031)
	[0.590]	[0.087]	[0.548]	[0.201]
Control Mean	.2908	.2229	.2996	.3484
N	2480	982	720	778
	(b) Analytic Sample			
	(1)	(2)	(3)	(4)
Treatment	0.007	0.049*	0.007	-0.044
	(0.018)	(0.026)	(0.035)	(0.032)
	[0.689]	[0.061]	[0.837]	[0.176]
Control Mean	.2985	.2257	.3203	.3535
N	2347	935	679	733

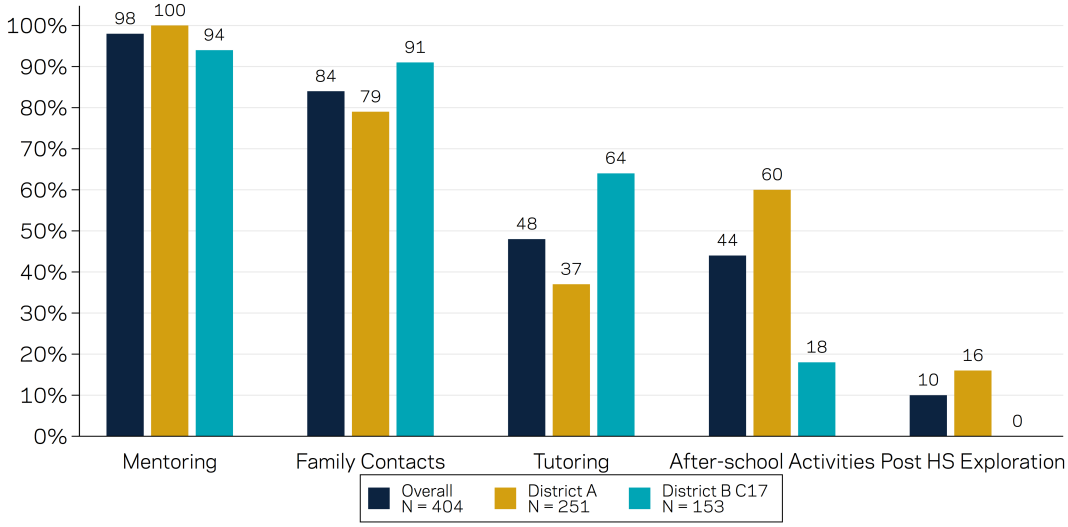
Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Ever offer is used as the treatment definition. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. All regressions include school-level and cohort fixed effects.

Table A.9: Counts for Each Activity, Program Enrollees

	Dist. A C17	Dist. B C17	Dist. B C18	Dist. A C17 - Dist. B C17	Dist. B C18 - Dist. B C17
	Means			Means (Standard Errors)	
	(a) Counts, Unconditional on Participation in Activity				
Tutoring Sessions	2.182	3.775	0.932	-1.59***	-2.84***
				(0.52)	(0.46)
After School Activities	4.673	2.999	2.668	1.67***	-0.33
				(0.55)	(0.52)
Post HS Exploration	0.245	0.0650	0.0472	0.18***	-0.018
				(0.040)	(0.019)
Mentoring Sessions	25.57	28.88	18.52	-3.31	-10.4***
				(2.74)	(2.71)
After School Activities	0.321	0.0825	0.0944	0.24***	0.012
				(0.046)	(0.029)
Employment Related Activities	0.0905	0.0595	0.0319	0.031	-0.028
				(0.026)	(0.027)
Contacting Family	3.293	5.785	3.616	-2.49***	-2.17***
				(0.55)	(0.55)
Student Enrichment	0.815	0.157	0.136	0.66***	-0.020
				(0.10)	(0.048)
Other/Unknown	3.405	2.677	2.401	0.73	-0.28
				(0.47)	(0.49)
Total Obs.	1,005	739	784		
	(b) Counts, Conditional on Participation in Activity				
Tutoring Sessions	12.75	19.38	7.615	-6.63***	-11.8***
				(2.20)	(1.90)
After School Activities	20.15	17.45	19.19	2.71	1.74
				(2.11)	(2.62)
Post HS Exploration	2.894	1.600	1.947	1.29***	0.35
				(0.39)	(0.37)
Mentoring Sessions	99.61	127.8	91.90	-28.2***	-35.9***
				(5.68)	(6.10)
After School Activities	3.167	2.033	2.242	1.13***	0.21
				(0.38)	(0.47)
Employment Related Activities	1.338	4	3.125	-2.66***	-0.87
				(0.95)	(1.45)
Contacting Family	13.62	25.75	17.94	-12.1***	-7.81***
				(1.47)	(1.35)
Student Enrichment	5.497	2.189	3.147	3.31***	0.96
				(0.63)	(0.73)
Other/Unknown	15.35	16.21	17.75	-0.87	1.54
				(1.96)	(2.55)
Total Obs.	1,005	739	784		

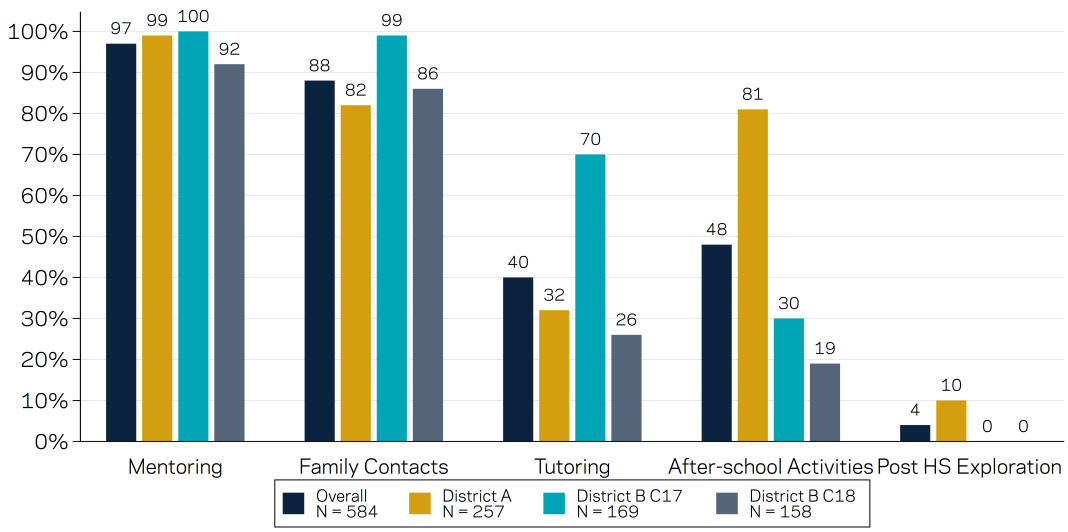
Notes: *** p<0.01 ** p<0.05 * p<0.1. Heteroskedasticity robust standard errors are in parenthesis. Sample includes students that enrolled into the program. Differences are not regression-adjusted.

Figure A.2: Share of students enrolled in the program with at least one record in a given activity category, 2017-18 school year.



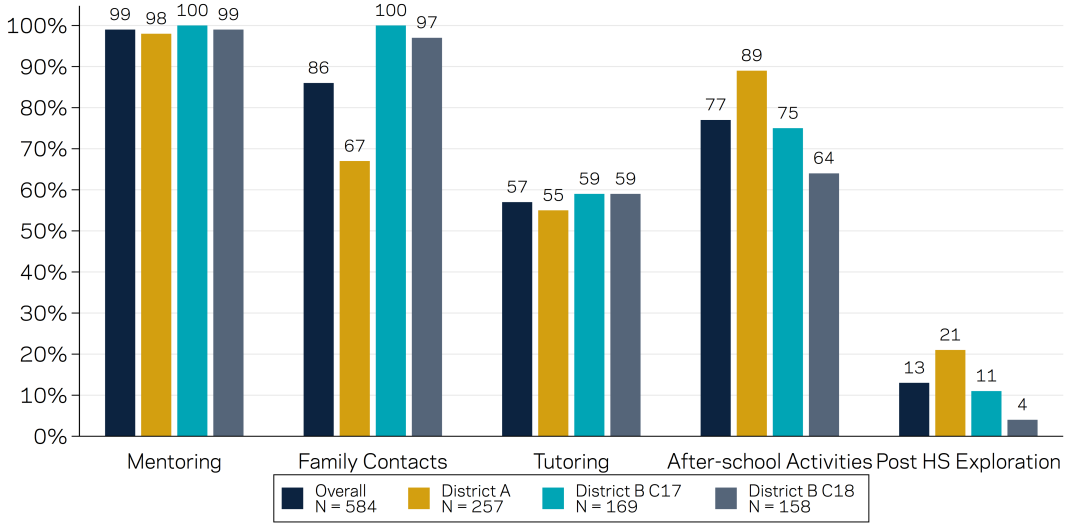
Note: Figure reports the share of students with at least one record in a given activity category during the 2017-18 academic year, which is freshman year for the 2017-18 cohort.

Figure A.3: Share of students enrolled in the program with at least one record in a given activity category, 2018-19 school year.



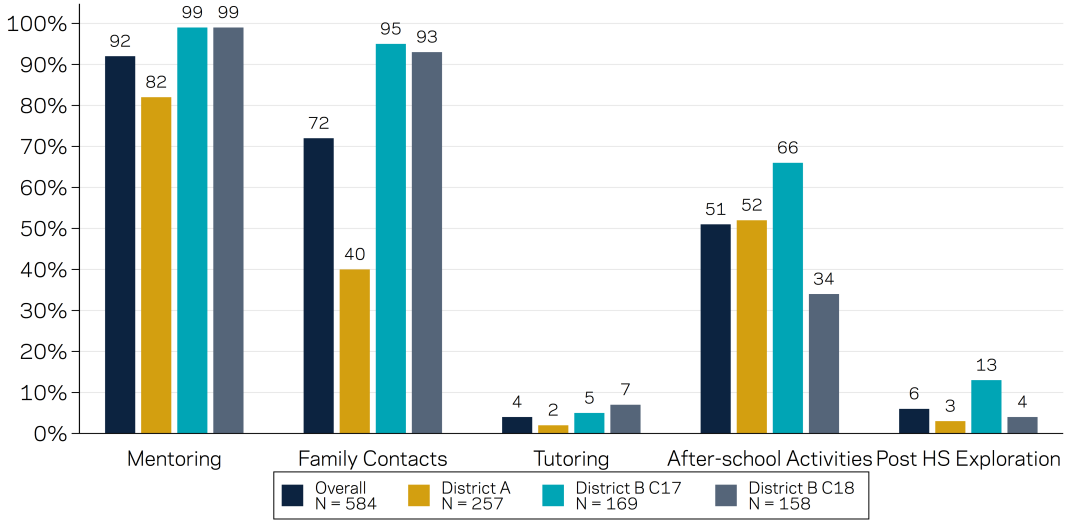
Note: Figure reports the share of students with at least one record in a given activity category during the 2018-19 academic year, which is sophomore year for the 2017-18 cohort and freshman year for the 2018-19 cohort.

Figure A.4: Share of students enrolled in the program with at least one record in a given activity category, 2019-20 school year.



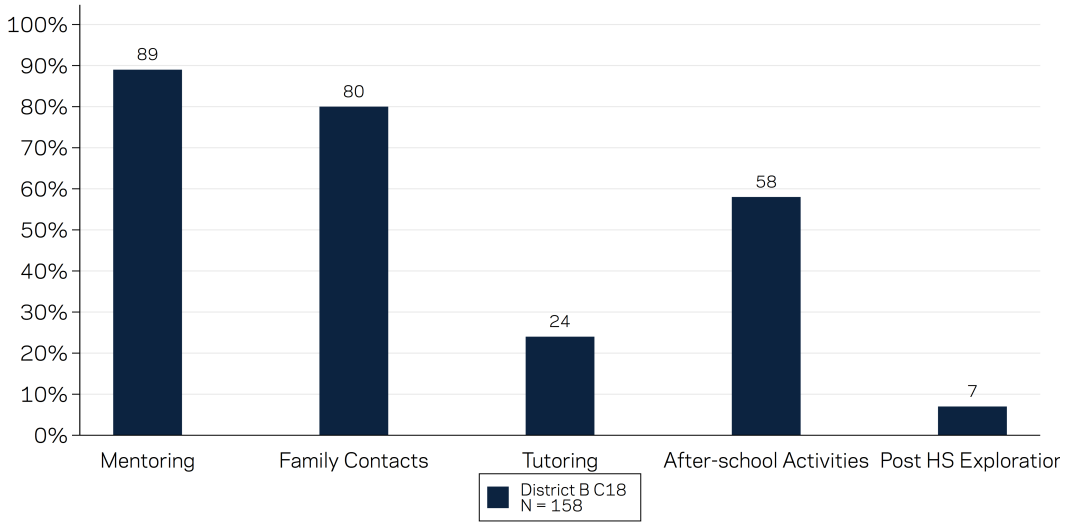
Note: Figure reports the share of students with at least one record in a given activity category during the 2019-20 academic year, which is junior year for the 2017-18 cohort and sophomore year for the 2018-19 cohort.

Figure A.5: Share of students enrolled in the program with at least one record in a given activity category, 2020-21 school year.



Note: Figure reports the share of students with at least one record in a given activity category during the 2020-21 academic year, which is senior year for the 2017-18 cohort and junior year for the 2018-19 cohort.

Figure A.6: Share of students enrolled in the program with at least one record in a given activity category, 2021-22 school year.



Note: Figure reports the share of students with at least one record in a given activity category during the 2021-22 academic year, which is senior year for the 2018-19 cohort.

Figure A.7: Core Service Dosage Goals

Activity	Prior 2019	2019 – 20	2020-21	2021-22	2022-23
Mentoring Contacts	• 1 successful contact per week for 90% of YA's caseload.	• 4 successful contacts per month for 85% of the YA's caseload Sept – June • 2 successful contacts per month in July and August for 85% of YA's caseload	• 4 successful or 3 successful and one attempted contacts per month for 80% of the YA's caseload Sept – June • 2 successful contacts per month July and August for 80% of YA's caseload	• 4 successful or 3 successful and one attempted contacts per month for 70% of the YA's caseload Sept – June • 2 successful contacts per month July and August for 70% of YA's caseload	• 4 successful or 3 successful and one attempted contacts per month for 80% of the YA's caseload Sept – June • 2 successful contacts per month July and August for 80% of YA's caseload
After/Out of School Activities	• Students participate in weekly enrichments.	• 75% of YA's caseload participate in 3 or more enrichments per month Oct – June	• 70% of YA's caseload participate in 2 or more enrichments per month Oct – June	• 70% of YA's caseload participate in 3 or more enrichments per month Oct – June	TBD
Academic Support Activities	• Students who's grade falls below a 70% (C) in a core subject attends after school tutoring.	• 50% of YA's caseload complete one or more academic support activities	• 50% of YA's caseload complete one or more academic support activities Oct - June	• 50% of YA's caseload complete one or more academic support activities Oct - June	TBD
YETA	• Students who meet AAA status become YETA certified.	• 85% of students who attend day 1 of YETA complete	• 70% of students who attend day 1 of YETA complete		TBD

Note: Table provides information on dosage goals for each student activity type and goal changes throughout the years of the study.

B Supplementary Tables

Table B.1: Baseline Balance Table by District and Cohort, Full Sample

	Control	Treatment	Total (District A)	Control	Treatment	Total (District B C17)	Control	Treatment	Total (District B C18)
	Means	Means	Treatment - Control	Means	Means	Treatment - Control	Means	Means	Treatment - Control
White	0.0962	0.0711	-0.025	0.228	0.204	-0.024	0.208	0.224	0.016
			(0.019)			(0.032)			(0.030)
Black	0.528	0.575	0.047	0.539	0.515	-0.024	0.493	0.502	0.0091
			(0.033)			(0.039)			(0.036)
Hispanic	0.321	0.293	-0.027	0.118	0.167	0.049*	0.141	0.138	-0.0030
			(0.031)			(0.026)			(0.025)
English Second Language	0.327	0.312	-0.015	0.209	0.202	-0.0072	0.276	0.243	-0.032
			(0.031)			(0.031)			(0.031)
In Poverty	0.950	0.949	-0.0037	0.898	0.909	0.012	0.935	0.946	0.011
			(0.014)			(0.023)			(0.017)
Enrollment Age (Years)	14.13	14.18	0.034	14.09	14.06	-0.028	14.11	14.01	-0.095**
			(0.039)			(0.049)			(0.042)
Predicted Graduation Rate	0.725	0.722	0.0025	0.737	0.737	-0.0027	0.728	0.725	-0.0055
			(0.0047)			(0.0064)			(0.0059)
Female	0.725	0.722	0.0049	0.737	0.737	0.0073	0.728	0.725	0.0066
			(0.033)			(0.039)			(0.035)
Over Two Risk Factors	0.725	0.722	0.025	0.737	0.737	-0.0069	0.728	0.725	-0.0056
			(0.033)			(0.039)			(0.035)
Total Obs.	344	661	1,005	254	485	739	355	429	784

Note: *** p<0.01, ** p<0.05, * p<0.10. Heteroskedasticity robust standard errors are used. Statistics are taken from the data provided by the school districts. The predicted graduation rate is a predicted value of the graduation rate based on the other covariates in the table. The number of observations for enrollment age and predicted graduation rate is one less for the District B sample due to missing information. Ever offer is used as the treatment definition.

Table B.2: Effect of Program on Credit Accumulation, Pooled and by District - Full Sample

	Pooled	District A	District B	
	ITT	ITT	ITTC17	ITTC18
	(a) Credits in Year 1			
	(1)	(2)	(3)	(4)
Treatment	-0.002	-0.006	0.406*	-0.266
	(0.112)	(0.166)	(0.214)	(0.205)
	[0.986]	[0.973]	[0.059]	[0.195]
Control Mean	4.866	5.427	4.272	4.749
N	2468	976	716	776
	(b) Credits Accumulated in Years 1 and 2			
	(1)	(2)	(3)	(4)
Treatment	0.646	0.646	0.000	0.000
	(0.399)	(0.399)	(.)	(.)
	[0.105]	[0.105]	[.]	[.]
Control Mean	16.79	16.79	0	0
N	852	852	739	783
	(c) Above Credit Threshold			
	(1)	(2)	(3)	(4)
Treatment	0.004	0.004	0.039	-0.051
	(0.029)	(0.029)	(0.038)	(0.034)
	[0.901]	[0.901]	[0.298]	[0.134]
Control Mean	.6091	.6091	.4897	.5556
N	976	976	716	776

Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Ever offer is used as the treatment definition. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. All regressions include school-level and cohort fixed effects. The sample is a subset of the 2,528 students in the full sample that had observable credit accumulation. Credit accumulation was only available in District A for panel (b) due to data issues in District B. Outcomes are conditional on enrollment in a school year for at least 10 days and attendance for at least one of those days.

Table B.3: Effect of Program on Chronic Absenteeism, Pooled and by District - Full Sample

	Pooled	District A	District B	
	ITT	ITT	ITTC17	ITTC18
	(a) Chronically Absent in Year 1			
	(1)	(2)	(3)	(4)
Treatment	-0.018	-0.013	-0.068**	0.020
	(0.018)	(0.030)	(0.034)	(0.031)
	[0.329]	[0.664]	[0.046]	[0.509]
Control Mean	.3601	.4565	.3725	.2606
N	2483	984	720	779
	(b) Any Chronic Absenteeism in Years 1 or 2			
	(1)	(2)	(3)	(4)
Treatment	-0.005	-0.024	-0.034	0.029
	(0.019)	(0.029)	(0.036)	(0.034)
	[0.787]	[0.415]	[0.348]	[0.394]
Control Mean	.5059	.6131	.4699	.4294
N	2498	990	726	782

Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Ever offer is used as the treatment definition. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. All regressions include school-level and cohort fixed effects. The sample is a subset of the 2,528 students in the full sample that had observable chronic absenteeism. Chronic absenteeism data is unavailable in District B Cohort 17 for year 3 and District B Cohort 18 for year 4 due to COVID-19. Both outcomes are conditional on enrollment in a school year for at least 10 days and attendance for at least one of those days.

Table B.4: Effect of Program on Taking and Passing Standardized Exams in First Year, Pooled and by District - Full Sample

	Pooled	District A	District B	
	ITT	ITT	ITTC17	ITTC18
	(a) Took Standardized Exam			
	(1)	(2)	(3)	(4)
Treatment	0.017	0.037	0.093***	-0.055*
	(0.018)	(0.029)	(0.031)	(0.030)
	[0.318]	[0.212]	[0.003]	[0.065]
Control Mean	.6264	.3837	.7323	.7859
N	2526	1004	739	783
	(b) Passed at Least 1 Standardized Exam			
	(1)	(2)	(3)	(4)
Treatment	0.022	0.030	0.080**	-0.035
	(0.018)	(0.028)	(0.035)	(0.032)
	[0.226]	[0.291]	[0.024]	[0.279]
Control Mean	.3757	.3256	.3543	.4394
N	2526	1004	739	783

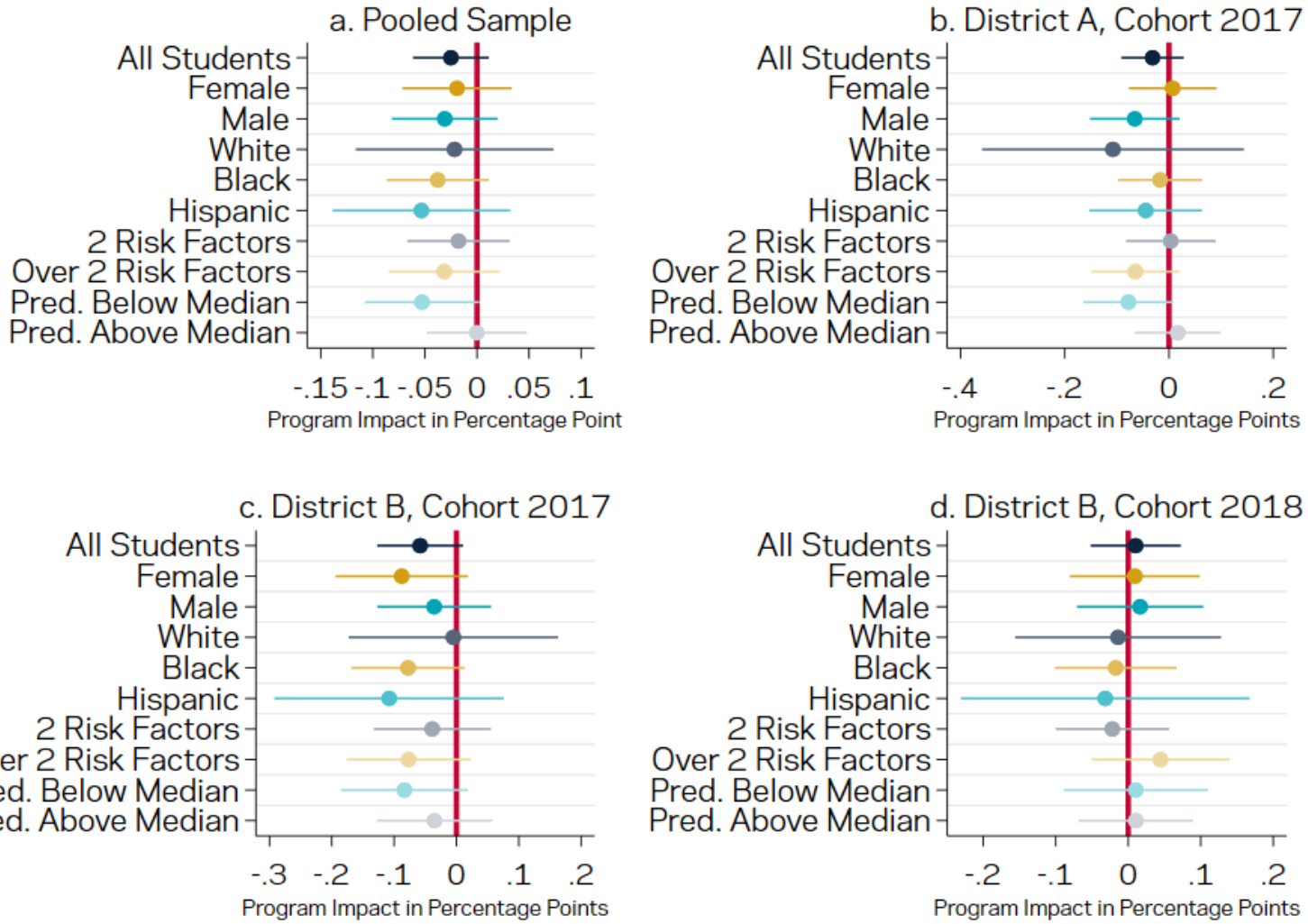
Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Ever offer is used as the treatment definition. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. The sample is a subset of the 2,528 students in the full sample that had observable standardized exam data.All regressions include school-level and cohort fixed effects.

Table B.5: Effect of Program on Post-Secondary Outcomes, Pooled and by District - Full Sample

	Pooled	District A	District B	
	ITT	ITT	ITTC17	ITTC18
	(a) Enrolled in Post-Secondary Education as of December 2024			
	(1)	(2)	(3)	(4)
Treatment	-0.066***	-0.052*	-0.031	-0.104***
	(0.019)	(0.029)	(0.036)	(0.033)
	[0.000]	[0.076]	[0.390]	[0.002]
Control Mean	.3827	.3353	.3898	.4225
N	2501	979	739	783
	(b) Total Semesters Enrolled as of December 2024			
	(1)	(2)	(3)	(4)
Treatment	0.390*	0.800**	0.372	-0.183
	(0.207)	(0.367)	(0.434)	(0.299)
	[0.059]	[0.030]	[0.393]	[0.542]
Control Mean	3.931	3.363	4.606	3.913
N	849	293	271	285
	(c) Completed Post-Secondary Education as of December 2024			
	(1)	(2)	(3)	(4)
Treatment	0.013*	0.016*	0.018	0.003
	(0.007)	(0.010)	(0.015)	(0.013)
	[0.067]	[0.086]	[0.236]	[0.824]
Control Mean	.02431	.01484	.0315	.02817
N	2501	979	739	783

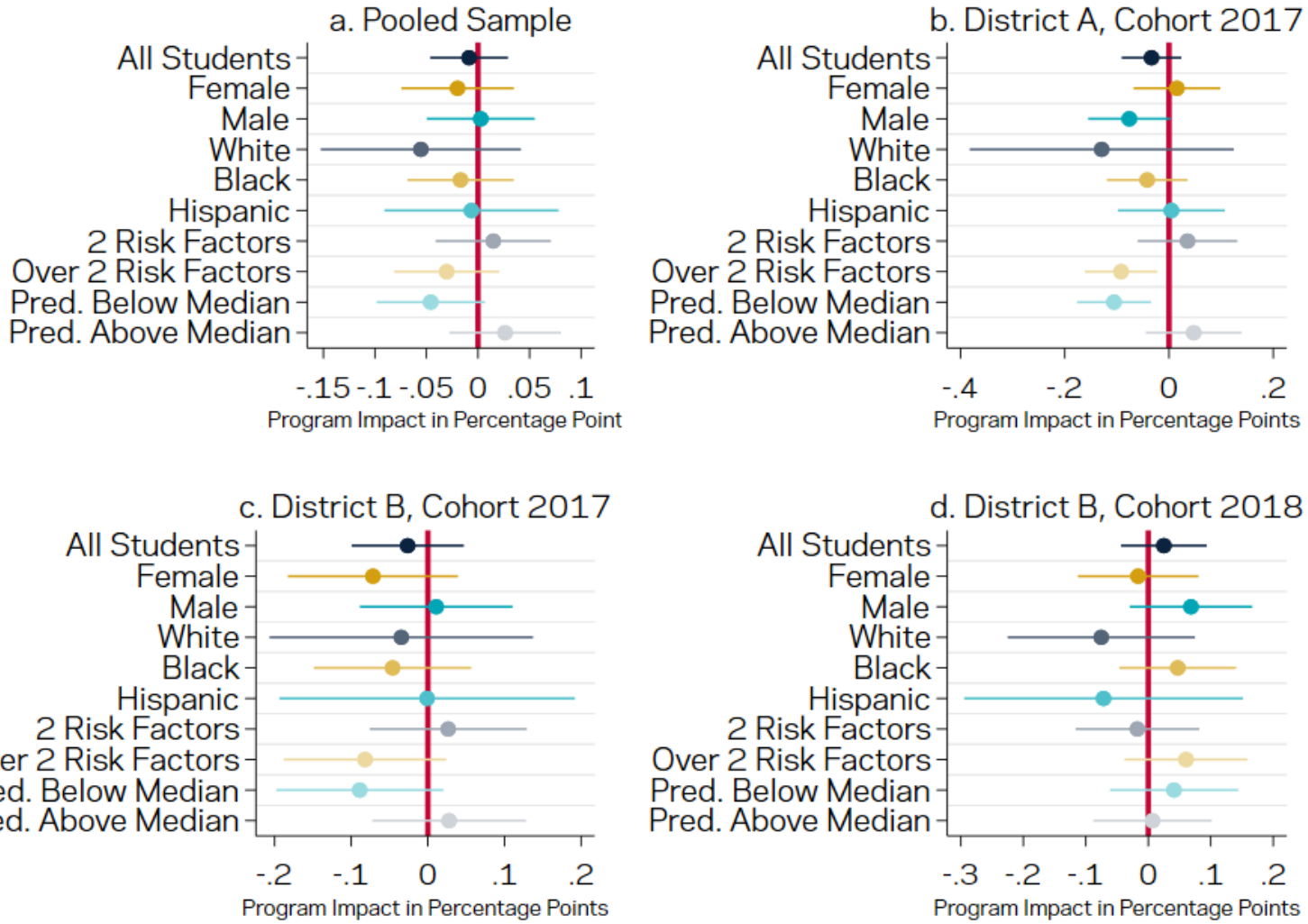
Notes: *** p<0.01 ** p<0.05 * p<0.1. Standard errors in parenthesis. p-values in brackets. Heteroskedasticity robust standard errors are used. Control variables include binary indicators for whether a student is female, black, white, Hispanic, spoke English as a Second Language, in poverty, age 15 at enrollment, age 16 or over at enrollment, and whether they have more than two risk factors. All regressions include school-level and cohort fixed effects. The sample is all of the students in our sample that also were able to match to National Student Clearinghouse data. Ever offer is used as the treatment definition. Total semesters enrolled is conditional on having enrolled into post-secondary education.

Figure B.1: Intent to Treat Estimates of Effect of Program on any Chronic Absence in Year 1



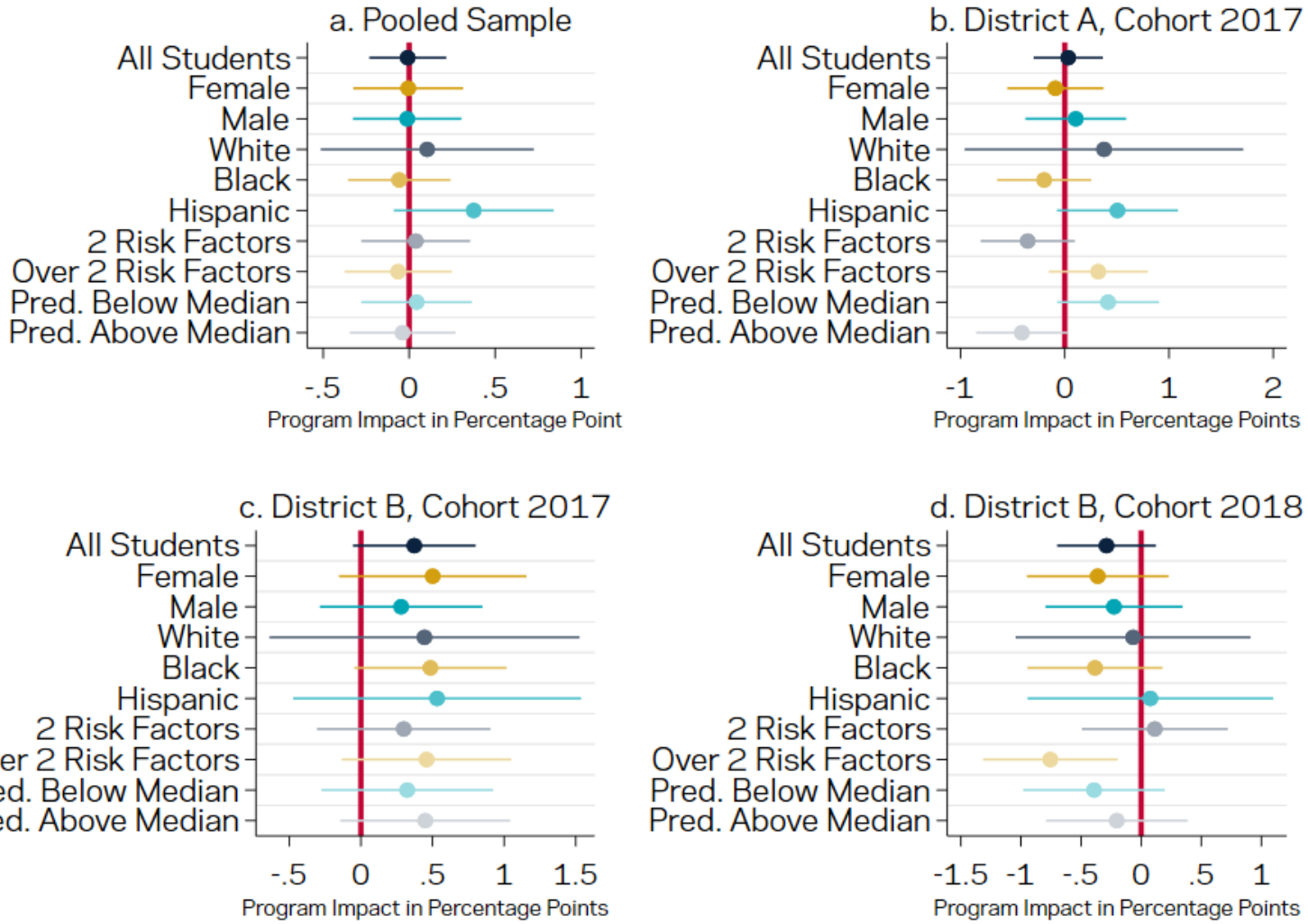
Notes: This figure uses ever offer as the treatment definition. All estimates include school-level and cohort fixed effects. Standard errors clustered by school/risk factor/gender/cohort. "Pred." represents "Predicted". The predicted below median bars represent the impact of the program for those who were predicted to be below the median graduation rate. The predicted above median bars represent the impact of the program for those who were predicted to be above the median graduation rate. The bars represent 95% confidence intervals.

Figure B.2: Intent to Treat Estimates of Effect of Program on any Chronic Absence in Years 1 or 2 for Subgroups



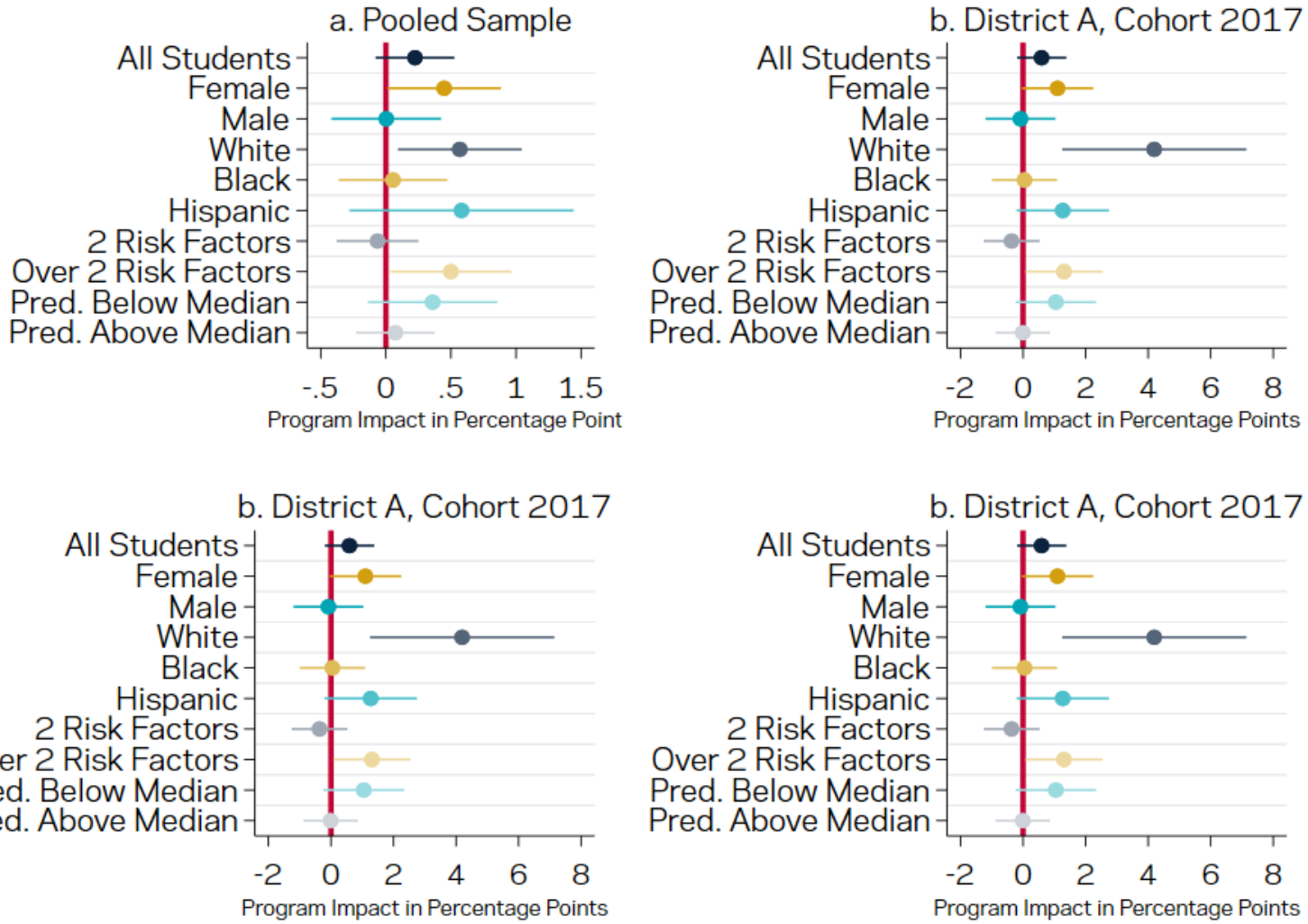
Notes: This figure uses ever offer as the treatment definition. All estimates include school-level and cohort fixed effects. Standard errors clustered by school/risk factor/gender/cohort. "Pred." represents "Predicted". The predicted below median bars represent the impact of the program for those who were predicted to be below the median graduation rate. The predicted above median bars represent the impact of the program for those who were predicted to be above the median graduation rate. The bars represent 95% confidence intervals.

Figure B.3: Intent to Treat Estimates of Effect of Program on Credits Achieved in First Year



Notes: This figure uses ever offer as the treatment definition. All estimates include school-level and cohort fixed effects. Standard errors clustered by school/risk factor/gender/cohort. "Pred." represents "Predicted". The predicted below median bars represent the impact of the program for those who were predicted to be below the median graduation rate. The predicted above median bars represent the impact of the program for those who were predicted to be above the median graduation rate. The bars represent 95% confidence intervals.

Figure B.4: Intent to Treat Estimates of Effect of Program on Cumulative Credits Achieved in First and Second Year



Notes: This figure uses ever offer as the treatment definition. All estimates include school-level and cohort fixed effects. Standard errors clustered by school/risk factor/gender/cohort. "Pred." represents "Predicted". The predicted below median bars represent the impact of the program for those who were predicted to be below the median graduation rate. The predicted above median bars represent the impact of the program for those who were predicted to be above the median graduation rate. The bars represent 95% confidence intervals.