# EFFECTS OF CROSS-AGE PEER MENTORING PROGRAM WITHIN A RANDOMIZED CONTROLLED TRIAL

Authors:
Eric Jenner, PhD[a]; Katherine Lass, MPH, LMSW[a]; Sarah Walsh, PhD[a]; Hilary Demby, MPH[a]; Rebekah Leger, MPH[a]; Gretchen Falk, MPH, RN[a]

[a] The Policy & Research Group, 8434 Oak Street, New Orleans, Louisiana 70118, USA; Phone: 504-865-1545; Fax: 504-865-0293

Email address for each author:
Eric Jenner: ejenner@policyandresearch.com
Katherine Lass: katie@policyandresearch.com
Sarah Walsh: sarah@policyandresearch.com
Hilary Demby: hilary@policyandresearch.com
Rebekah Leger: rebekah@policyandresearch.com
Gretchen Falk: gretchen@policyandresearch.com

Corresponding author: Katherine Lass, 8434 Oak Street, New Orleans, Louisiana, 70118; katie@policyandresearch.com; Phone: 504-865-1545; Fax: 504-865-0293

**Funding**
This work was supported by the U.S. Department of Education (ED), Office of Elementary and Secondary Education (OESE), Investing in Innovation (i3) Grant U411C150048. The i3 program required a rigorous design for studies implemented under this funding, and provided technical assistance through Abt Associates for decisions around study design and analysis; however, OESE was not involved in designing the study or performing analysis of the data. Additionally, all data collection activities were carried out independently by research staff.

**Conflict of interest**
No potential competing interest to report.

**Acknowledgments**
This study would not have been possible without the following field team members at PRG: Kelly Burgess, Elyse Mason, Teresa Smith, Jordan Stribling, Charlee Roundhill, Carolyn Kelly, Alexis Mayfield, and Noor Qaragholi. We would also like to thank Bruce Randel, PhD, from Century Analytics for his thoughtful guidance and interminable encouragement throughout the life of this study. In addition, we are thankful for the work of staff at the Center for Supportive Schools who both developed the intervention and worked diligently throughout the study to provide sites with training and technical assistance in program implementation. Finally, we express deep appreciation to the school administrators, staff, and students who participated in this study.

# ABSTRACT

This paper summarizes results from an impact study of a cross-age peer mentor program designed to prevent school dropout during the transition from middle to high school. The study employed a randomized controlled trial (RCT) of 1,351 ninth-grade students. Although the intent-to-treat (ITT) analyses indicate modest, yet potentially meaningful program impact on ninth-grade outcomes of disciplinary action, school attachment, and expectations of degree attainment across varying dosage levels, complier average causal effect (CACE) estimates suggest that a threshold level of program participation broadens the program's impact on additional academic achievement and social and emotional learning outcomes. Given the adverse effects of the transition to high school, this promising evidence indicates that the cross-age peer mentoring intervention could be a cost-effective and sustainable strategy for high schools to implement. In presenting the study findings, we outline two methods for estimating the CACE, or the effect of program participation, that allow researchers to leverage RCT data, beyond ITT analyses, to provide meaningful context for how and when interventions work.



## KEYWORDS

Complier average causal effect (CACE), student engagement, high school transition, instrumental variable regression, principal score, peer mentoring, Randomized Control Trial (RCT), intent to treat, academic achievement, social and emotional learning

# INTRODUCTION

The National Center for Education Statistics reported that in 2018, there were 2.1 million youth between the ages of 16 and 24 who were either not enrolled in school or who had not earned a high school degree (Hussar et al., 2020). Failure to obtain a high school degree is associated with increased use of public assistance and increased probability of incarceration; further, research suggests that a five percent increase in male high school graduation rates could add up to $1.2 billion in additional annual earnings to enter the national economy (Alliance for Excellent Education, 2013; Cohen & Smerdon, 2009). The transition to ninth grade, sometimes referred to as the "9th grade shock," is frequently associated with sharp declines in academic performance and increases in absenteeism and discipline issues (Cohen & Smerdon, 2009; Pharris-Ciurej, Hirschman, & Willhoft, 2012; Smith, 2006). Research demonstrates that how a student performs in ninth grade is highly predictive of high school graduation rates (Allensworth & Easton, 2005; Easton, Johnson, & Sartain, 2017).

The U.S. Department of Education's Investing in Innovation (i3) Fund provided grants from 2010 to 2016 for the purpose of implementing and rigorously evaluating innovative educational practices that promote student achievement, close achievement gaps, decrease dropout rates, and increase high school graduation and postsecondary enrollment rates (U.S. Department of Education, 2015). Through a five-year i3 grant received in 2016, Peer Group Connection-High School (PGC-HS), a school-based, cross-age peer mentoring program designed to ease the transition from middle to high school for ninth-grade students, was implemented in six public high schools in rural North Carolina during the 2016–2017, 2017–2018, and 2018–2019 school years. PGC-HS aims to mitigate declines in academic performance and student engagement in ninth-grade students by connecting them with junior and senior peer mentors, who create a supportive environment where students work through a structured curriculum of group activities that address the unique challenges associated with this transition.

In this study, we report findings from an individual-level randomized controlled trial (RCT) that assessed the efficacy of the PGC-HS program on ninth-grade outcomes related to dropout prevention. An RCT design is the gold standard approach to assess cause-effect relationships of interventions, and is commonly used in educational evaluations to assess program impacts on outcomes of interest, including academic achievement, attainment, persistence, and retention. Typically, RCT evaluations employ intent-to-treat (ITT) analyses to answer the question of whether the program was effective for students who were randomly assigned to the intervention under study. However, this effect may not always be the only or the primary estimate of interest – particularly when the intervention has a complex, multi-session structure (Stuart, Perry, Le, & Ialongo, 2008). For these interventions, the ITT estimate often offers an incomplete picture of the efficacy of the intervention. Any variation in the post-randomization experience of the program is conflated by the ITT average treatment effect (ATE) estimate, and this makes it difficult for program developers and consumers to understand whether the program was effective for students who received the program as it was intended (i.e., at specific levels of dosage). To answer this question, the evaluator must turn to alternative causal estimates. One such estimate is the complier average causal effect (CACE), which estimates the effect of the program at a specified level of dosage. In this study, we employ both frameworks – the ITT and the CACE – to estimate the efficacy of PGC-HS on ninth-grade student outcomes related to dropout prevention. Specifically, we attempt to answer the following two research questions:

1. ITT analysis: What is the ATE of the offer to participate in a cross-age peer mentoring program relative to the offer to receive the control condition (class as usual) on ninth-grade participants' academic, behavioral, and noncognitive outcomes related to dropout prevention?

2. CACE analysis: What is the effect of participating (in 14 or more sessions) in a cross-age peer mentoring program relative to the control condition (class as usual) on ninth-grade participants' academic, behavioral, and noncognitive outcomes related to dropout prevention?

## ENGAGEMENT IN HIGH SCHOOL

Ninth grade is a challenging year for students. Ninth-grade students demonstrate higher rates of course failure, declines in test scores, and behavioral problems more than any other grade. Research has found that how a student performs in ninth grade is directly connected to the probability of high school graduation (Allensworth & Easton, 2005). For example, the On-Track indicator from the University of Chicago identifies a ninth-grade student as being on track if they earn at least five full-year course credits and one or fewer failing grades in a core course (Allensworth & Easton, 2005). Those students who were on track by the end of ninth grade were more than 3.5 times more likely to graduate in four years than students defined as off-track. The barrier of ninth grade is also evident in enrollment figures. Sometimes referred to as the "ninth-grade bulge," ninth grade has a higher enrollment rate than any other grade in high school, due in part to high retention rates (Keaton, 2012; Pharris-Ciurej et al., 2012).

Maintaining a high level of engagement in ninth grade may improve the likelihood that a student stays on track to graduate in four years. Increased engagement is associated with positive academic outcomes and a decreased likelihood of dropping out (Christenson, Reschly, & Wylie, 2012; Fredricks, Blumenfeld, & Paris, 2004). Although specific definitions of engagement vary widely, researchers tend to agree that there are at least three dimensions of engagement: behavioral, emotional, and cognitive (Appleton, Christenson, & Furlong, 2008; Fredricks et al., 2004). Other schools of thought posit academic engagement as a fourth dimension (Christenson et al., 2012). Behavioral engagement includes positive participation in school-related activities, including behaviors such as attendance, effort, concentration, and persistence. Emotional engagement represents the affective reactions, both positive and negative, to teachers, classmates, schoolwork, and the physical school environment. Cognitive engagement is viewed as both motivation and the personal investment in the learning process, or the effort necessary to complete complicated tasks and understand complex ideas. Students who are cognitively engaged self-regulate their learning strategies. Engagement is presumed to be malleable and vary in intensity and duration depending on context (Fredricks et al., 2004). The impressionable nature of engagement and its association with sustained academic achievement and retention suggest that an intervention that aims to strengthen a student's behavioral, cognitive, and emotional engagement early in the high school experience could increase the likelihood that they will remain on track and invested in their potential to graduate.

## PEER MENTORING AS PREVENTION

Mentoring has become a popular strategy to prevent students from becoming disengaged and dropping out, and has been shown to help students stay in and progress through school (U.S. Department of Education, 2017). Research with adolescents has demonstrated that connections to others are important for success, and that adolescents with natural mentors are found to be more likely to complete high school and college, have higher self-esteem and life satisfaction, and be less likely to use illicit substances or be involved in nonviolent delinquency (Bernat & Resnick, 2009; Zimmerman, Bingenheimer, & Notaro, 2002). In particular for ninth graders, one study found that ninth-grade students who were at high risk of dropping out and who participated in a school-based mentoring program earned more credits by the end of ninth grade as compared to students who had not gone through the programming (Chan, Kuperminc, Seitz, Wilson, & Khatib, 2020). Some research finds that

mentoring programs have the largest effect on adolescents from disadvantaged backgrounds, though this may be due to the fact that mentoring programs are typically developed for students at higher risk of poor outcomes (Bernat & Resnick, 2009).

Students are more likely to succeed academically when they feel connected to their school, and adolescents' educational outcomes are heavily influenced by the behaviors and attitudes of their peers (Centers for Disease Control and Prevention [CDC], 2009). As a result, the CDC recommends positive peer support groups and peer mentoring strategies as a means to help reinforce positive behaviors such as empathy, stress management, and conflict resolution among students who are at risk academically. Cross-age peer mentoring aims to build engagement and promote academic success by giving at-risk students the opportunity to form positive personal relationships with older students who have persisted through similar experiences and who act as positive role models.

## PEER GROUP CONNECTION – HIGH SCHOOL

This study examines the impact of the PGC-HS program on academic, behavioral, and noncognitive outcomes indicative of school engagement and academic achievement. PGC-HS is a school-based high school transition and cross-age peer mentoring program for ninth-grade students that is designed to improve their engagement in school and educational outcomes. The program leverages existing resources, such as school staff, parents, and student leaders, to create a supportive environment for transitioning high school students that encourages them to attend school, set personal goals, work hard, and make healthy decisions. By offering additional support to ninth-grade students, the program seeks to mitigate the problems associated with the transition to high school, such as disengagement from school, absenteeism, increases in disciplinary events, and declines in academic performance.

PGC-HS is designed to foster supportive relationships between students and strengthen their connection to the broader school community to improve social and emotional skills, engagement, persistence, and retention. Older students (juniors and seniors) are recruited to create a nurturing and supportive environment for incoming ninth graders. A stakeholder team of administrators, faculty, parents/caregivers, and community members support program implementation and advise two faculty advisors who run the program and teach a daily leadership course to the junior and senior peer leaders, who earn credit toward graduation. Peer leaders, who serve as role models, discussion leaders, and mentors, work in pairs to co-lead diverse groups of 10 to 12 ninth graders in weekly outreach sessions that follow a structured but flexible curriculum. By participating in the weekly outreach sessions, ninth-grade students practice academic, social, and emotional skills such as critical thinking, goal setting, decision-making, conflict resolution, teamwork, and communication. The ninth-grade students not only have access to older student mentors, but they form relationships with a diverse group of other ninth graders. The program is designed to be implemented with regularity throughout the fall semester or across the full school year, depending on individual school scheduling capacity.

## ESTIMATES OF IMPACT AND COMPLEX INTERVENTIONS

A well-executed RCT is considered the highest level of scientific evidence because it permits a causal estimate of a treatment effect of an intervention with a minimum of bias. What that treatment is, however, depends on the analytical framework employed by the researcher. The benchmark is the ITT framework, which assesses the impact of the offer to participate in an intervention and not the exposure to the intervention itself. The ITT estimate sidesteps post-randomization selection bias by including all participants with outcome data in the analytic sample, regardless of actual exposure. Notwithstanding this apparent insensitivity, the ITT is generally the preferred impact estimate of policy

stakeholders because it incorporates varying uptake and compliance – dilutive features that will undoubtedly exist in the real world beyond the efficacy study – into the estimate of impact. The trade-off, however, is that by ignoring dosage/compliance and quality of implementation as a consideration, the ITT estimate will understate the efficacy of an intervention as it was intended to be delivered and may not be replicable given different implementation contexts. Given that many program evaluations, particularly in education research, are derived from large-scale, multisite RCTs, which can be expensive and time-consuming, it is increasingly apparent that some estimate of the efficacy – or the impact of the intervention when some threshold of quality and/or dosage is realized – is equally relevant.

This is especially true for efficacy trials of innovative practices and complex multi-session interventions, where noncompliance may be ameliorated by relatively easy-to-fix features such as implementation fidelity, facilitator training and practice, or encouraging individual-level uptake. As a result, researchers and consumers of research often want to know what it is about the program (quality of implementation, intensity of dosage, particular components) that is driving the average impact estimate. Indeed, an entire issue of the *American Journal of Evaluation* has been dedicated to unpacking the "black box" of ITT estimates while another issue of *New Directions for Evaluation* discusses design and analytic strategies for doing so within social and behavioral intervention research (Peck, 2016; Rallis, 2015). Moreover, Stuart et al. (2008) contend that the effect of complying with a program may be more generalizable than the ITT effect as rates of compliance may vary across different populations and within different studies, but the effect of full participation in the program should not necessarily vary in the same way. Moulton, Peck, & Greeney (2018) argue that program managers and funders with finite budgets can use CACE findings to help them make decisions about how to efficiently spend resources and effectively balance program intensity or duration with the number of participants served.

There are two pathways through which researchers can isolate causal effects of dosage in complex, multi-session interventions. The first is through prespecified and exogenous variation in the design stage. Bell and Peck (2016) offer three classic examples: (1) a multi-arm experimental trial where a control group is compared with two or more treatment groups that experience an increasing number of programmatic components; (2) a multistage, sequential trial that involves multiple points of random assignment as individuals move through the different stages of an intervention; (3) and a factorial design that crosses two program features to form a matrix of treatment groups. Each of these approaches requires a larger sample than a traditional two-arm randomized trial, which can substantially increase the complexity and cost. When these choices are not justifiable or for whatever reason are not included as exogenous features in the research design, researchers can turn to post hoc, quasi-experimental analytic strategies to estimate dosage-sensitive or compliance effects of multi-session interventions; methods include instrumental variable (Angrist, Imbens, & Rubin, 1996), propensity score (Rosenbaum & Rubin, 1983), and analysis of symmetrically predicted endogenous subgroups (ASPES) (Peck, 2003).

It is with this latter quasi-experimental pathway that we supplement the ATE estimates produced by an ITT analysis. Specifically, we aim to assess the effect of the PGC-HS program when students participate in, or comply with, the full program using two analytic methods derived from the principal stratification framework that estimates impact conditioned on endogenous (post-randomization) compliance (Frangakis & Rubin, 2002).

# METHODS

This study is part of a multisite evaluation, funded by a five-year i3 grant, that employed an individual-level RCT to determine whether the PGC-HS program improves academic, behavioral, and noncognitive outcomes indicative of school engagement and academic achievement.

## PARTICIPANTS/SETTING

A total of 1,532 ninth-grade students from six public high schools in rural North Carolina were initially enrolled in the RCT during the 2016–2017, 2017–2018, and 2018–2019 school years. Three schools implemented PGC-HS during two school years and contributed two cohorts of study participants, for a total of nine distinct blocks. All sites were located in Local Education Associations (LEAs) eligible for the Rural and Low-Income Schools program, such that at least 20% of children aged 5 to 17 served by the LEA are from families with incomes below the poverty line. To be eligible for the study, students had to be entering ninth grade for the first time, be enrolled at a study school at the time of randomization, and be able to complete a participant questionnaire unassisted in either English or Spanish within 60 minutes.

Students from one high school (which included both PGC-HS and control students) were ultimately dropped from the RCT because the school experienced significant challenges implementing the PGC-HS intervention during the fall 2018 semester as a result of the impact and aftermath of Hurricane Florence in eastern North Carolina in August 2018. Because students were individually randomly assigned within schools and the hurricane event impacted students in both PGC-HS and control conditions in the same manner, the decision to drop these students from the RCT does not compromise the randomized design and we do not count these 181 students toward study sample attrition calculations.[1] Descriptive characteristics of the 1,351 ninth-grade students included in the RCT, both PGC-HS and control, are provided in Table 1.

[Table 1 goes here]

## PROCEDURES

Eligible students who did not opt out were individually randomly assigned to be offered the PGC-HS program (treatment) or classes as usual (control). Blocking was done at the school and cohort level. Three schools participated in the study for two years and contributed two cohorts of ninth-grade participants for a total of eight distinct blocks. For the remainder of this report, we refer to these distinct blocks as the *eight participating schools*. Each year, at the end of the first two weeks of the fall semester, school officials sent the research team a roster of ninth graders who were enrolled at the school and eligible for the study. Researchers then produced individually randomized rosters, blocked by study site and cohort year, using the random allocation (*ralloc*) command in Stata 14 (StataCorp, 2015). The assignment ratio was 1:1 treatment to control. Students were considered enrolled in the ITT sample at the point of random assignment. Randomized rosters were sent back to the school official, who then placed the students assigned to the treatment condition into the PGC-HS program, formed each of the peer groups, and assigned peer leader pairs to each group. Students who entered school after the point of randomization may have been placed into the PGC-HS program but were not enrolled in the study. Randomized students who later transferred or dropped out of the school remained in the ITT sample. Students in the ITT analytic sample remained in their randomized treatment condition, regardless of

[1] As explained in more detail below, when these 181 students are counted toward attrition calculations, the analytic samples for each outcome still meet the low-attrition threshold under both cautious and optimistic assumptions established by the What Works Clearinghouse.

program exposure. Fidelity to treatment assignment was managed by the school but monitored by the researchers. Students who were assigned to the control condition remained in their usual class(es) during the time when PGC-HS students were pulled from class to meet with their peer groups for either once/week 45-minute or twice/week 30-minute outreach sessions.

## DATA COLLECTION

Data were collected from two sources. Student baseline and ninth-grade administrative records compiled by study school data managers were sent directly to the evaluators and contained background demographic data, academic and administrative data (attendance, credits earned, GPA), and behavioral outcome data (suspensions, detentions, disciplinary referrals). Study participants completed a participant questionnaire twice – once at the beginning of the fall semester of their ninth-grade year (pre-program) and again during the spring semester when PGC-HS programming had concluded at their school (post-program). The questionnaire was comprised of measures of noncognitive outcomes (e.g., school attachment, growth mindset, educational ambitions, etc.). The research team administered the questionnaires in person at study schools with small classrooms of participating students. Methods of data collection were identical for treatment and control students.

## OUTCOMES

Outcomes were assessed using either administrative records or self-reported data collected at the end of the regular ninth-grade year. We examine the following academic and behavioral outcomes, as reported in a student's administrative record(s):

- *Credits:* a count variable indicating the total number of credits earned toward graduation during the regular school year;
- *GPA:* a continuous variable indicating the student's cumulative weighted GPA as of the end of the regular school year;
- *Attendance:* a count variable indicating the total number of days in attendance at school during the regular school year;
- *Promoted:* a dichotomous variable that indicates if the student was promoted to 10$^{th}$ grade at the end of the regular school year (1) or retained (0);
- *Suspension:* a dichotomous variable indicating whether (1) or not (0) a student received one or more suspensions during the regular school year;
- *Detention:* a dichotomous variable indicating whether (1) or not (0) a student received one or more detentions during the regular school year;
- *Disciplinary Referral:* a dichotomous variable indicating whether (1) or not (0) a student received one or more disciplinary referrals during the regular school year.

We also examine the following 11 noncognitive outcomes indicative of engagement, social emotional learning skills, and educational outlook, gathered via self-report questionnaires:

- *perceived self-efficacy in goal setting skills, growth mindset, grit, school attachment, peer connection, educational ambitions, educational expectations, educational aspirations, decision-making skills, peer norms*, and *social competence.*

Nine of the noncognitive outcomes were operationalized as mean scale scores from questionnaire items with 7-point Likert-type or semantic differential response scales. Scale scores were constructed by estimating the mean of all items that made up the scale and were only estimated if a student responded

to all items within a specified scale.[2] We did not impute any missing values in these or any outcome measures. Two noncognitive outcomes (educational aspirations and educational expectations) were constructed as dichotomous indicators of whether a student wanted to (aspiration) or thought they would (expectation) obtain a four-year degree or higher.

We predefined attendance and credits earned as the confirmatory outcomes for the impact evaluation. Attendance and credits earned are defined in the What Works Clearinghouse Dropout Prevention Review Protocol (U.S. Department of Education, 2014) as acceptable measures of staying in and progressing through school, respectively. As they are in separate domains, we do not employ multiple comparison correction. In addition to these, we also explore a broader set of outcomes that are central to the program's theory of change. Because academic success and social connectedness are important factors for improving student persistence and retention, the program's activities emphasize the development of critical thinking skills, healthy decision-making skills, a sense of belonging, positive peer relationships, and motivation to attend and do well in school. We include these outcomes because they are a part of the program's theory of change and also important academic markers and student-reported skills and attitudes that are associated with increased engagement and thereby academic achievement in school.

## ANALYTIC METHODS

### *RESEARCH QUESTION 1 – ITT ATE ESTIMATE*

We assess average impacts of assignment to the PGC-HS program using a regression equation that models the outcome of interest as a function of treatment status, and a series of covariates, including the baseline measure of the outcome variable (or a proxy if unavailable), age at baseline, race/ethnicity, gender, Individualized Education Plan (IEP) and English Language Learner (ELL) statuses at beginning of ninth grade, and a series of dummy variables representing randomization blocks. We estimate ATE effects with an Ordinary Least Squares (OLS) Regression/Linear Probability Model (LPM) for ease of interpretation. In the case of binary outcomes, we test the robustness of this approach with a logistic regression. No substantive differences were observed as a result of these analytic decisions.

### *RESEARCH QUESTION 2 – CACE ESTIMATE*

We assess the CACE by first defining full participation, assigning compliance/noncompliance status, and then using two analytic approaches: (1) principal score analysis using a weighted OLS/LPM regression, and (2) an instrumental variable analysis using two-stage least squares regression.

#### COMPLIANCE

Full participation in the PGC-HS program is defined as having attended at least 14 outreach sessions. The complete PGC-HS curriculum includes 26 standard outreach sessions and a number of supplemental sessions such as a Family Night and service-learning days. Schools are advised to hold a minimum of 18 outreach events in order to meet the minimum threshold of implementation fidelity, with flexibility in the types of sessions held. A moderately stringent definition of compliance was adopted because regular and persistent attendance at outreach sessions fostering the development of interpersonal relationships with peers and peer leaders is at the core of the intervention's logic model. We defined full participation as attending 14 or more outreach sessions; 61% of treatment students met this threshold.

[2] Scale reliability statistics can be found in Table S5 in the Supplementary Materials.

In the principal stratification framework, noncompliance/compliance is conceived of as a fixed but unknown individual-level propensity at baseline that can be partially observed in the treatment group and, with a few strong assumptions, be identified in the comparison group. The identification strategy is then used to estimate an unbiased treatment effect within that stratum (e.g., the CACE). For students who attend schools that fail to offer the minimum participation threshold of 14 outreach sessions, however, individual-level compliance is not observable at all (i.e., in the treatment group). In these schools, no student complies with treatment, but, and importantly, the noncompliance is not motivated by an individual's fixed propensity but by school-level features. These schools are excluded from CACE estimates because there is no within-school variation (IV regression) or because the stratum of compliers could not be explicitly identified (principal scores).

Of the eight participating school blocks, six offered at least 14 sessions thereby permitting students the opportunity to meet the compliance threshold; two did not.[3] Schools differed greatly in the number of PGC-HS sessions they offered to students, depending on the length of time they offered the program (fall semester only or the full academic year), the frequency with which they scheduled outreaches during that time (weekly 45-minute sessions or twice-weekly 30-minute sessions), and the prevalence of other community-level barriers to implementation. The school that offered the fewest number of sessions held 13, whereas the school that offered the most held 32 sessions.[4]

Below we describe the two methods used to estimate the effect of participating in the PGC-HS program: principal score weighting and two-stage least squares regression.

### PRINCIPAL SCORE METHOD

The first method we use to estimate the CACE is a balancing procedure, based on propensity score methods, which can be used in settings where principal stratum membership (compliance) is observable under one treatment condition. The use of propensity scores (Rosenbaum & Rubin, 1983) is a common technique to balance treatment and comparison groups in nonexperimental studies; however, propensity score methods have also been used to address noncompliance in RCTs as well (Follmann, 2000; Hill, Waldfogel, & Brooks-Gunn, 2002, 2003; Jo & Stuart, 2009; Stuart & Jo, 2015). Within the context of a randomized trial using a 1:1 assignment ratio, there is an expectation that principal stratum membership should be equally allocated to the treatment and control groups. In other words, if there are compliers in the treatment group, we would expect there exists a similar group of individuals in the control group who would have complied if the program had been offered to them. Whereas the conventional use of propensity scores aims to model treatment group membership (where treatment group membership is the same as intervention receipt), the aim here is to use propensity scores to model treatment receipt (compliance) in the treatment group and subsequently predict probability of principal stratum membership among members of the control group. In accordance with Hill et al. (2002), to distinguish this latter prediction step for the control group, we refer to their probability of principal stratum membership as the principal score.

The core assumption in propensity score methods is that of conditional ignorable treatment assignment, or the assertion that treatment assignment is independent of the potential outcomes, given a set of

[3] As a reminder, a ninth block of students was initially randomized to be included in the study, but was ultimately excluded from the RCT because of the impact of Hurricane Florence. The shortened fall semester as a result of a two-week closure and displacement of students in the school community prevented the school from implementing the PGC-HS program as it intended.

[4] We offer additional context on the average impact of the PGC-HS program when it is implemented with high fidelity (i.e., the school offered 18 or more sessions) in a table of results of an ATE analysis on outcomes when we exclude these two additional locations from the analytic sample (Table S6 of the Supplementary Materials).

observed covariates (Rosenbaum & Rubin, 1983). When we use a probability score to balance treatment and control groups to estimate CACE in an RCT, this assumption now applies to principal stratum membership (compliance). In other words, principal stratum membership is independent of the potential outcomes given the observed set of covariates (Jo & Stuart, 2009).

We follow the steps outlined in Stuart and Jo (2015) to estimate the CACE using principal score weights. Briefly, these include (1) using baseline covariates to predict compliance among the treatment group; (2) predicting probability of compliance (principal score) among members of the control group; (3) creating analytic weights reflecting probability of compliance; and (4) estimating CACE by fitting the outcome model using the principal score weights. Consistent with the ITT analysis, the outcome model was fit with OLS and included the following covariates: age, race/ethnicity, gender, IEP status, ELL status, randomization blocks, and the baseline measure of the outcome (or a proxy).

As with all propensity score methods, we are unable to directly test the ignorability assumption. In the context of an RCT, however, we do have the advantage of knowing that randomization should (in expectation) balance stratum membership in both groups. Although we can never know if our weighting strategy is sufficient, we have a reasonably robust collection of baseline covariates to predict both compliance and the outcomes. To assess the degree to which our treatment compliers and weighted control group are similar on observed covariates, we calculate the standardized differences in covariate means and proportions by running the same model described in the Baseline Equivalence section below, but with the analytic weights (generated in Step 3) added to the model.

### INSTRUMENTAL VARIABLE METHOD

The second method is an instrumental variable approach that uses the random assignment mechanism to act as an instrument for compliance to estimate the CACE. We estimate the CACE with a joint model that first estimates participation, given treatment assignment and subsequently estimates the outcome, given participation; this is known as Two-Stage Least Squares (TSLS) regression (Angrist & Imbens, 1995). Instrumental variable analysis is a common technique in evaluation to estimate the CACE in randomized trials (Black et al., 2006; Connell, 2009; Dunn et al., 2003; Schochet & Chiang, 2011; Stuart et al., 2008). The instrumental variable approach relies on five key assumptions that are necessary for causal interpretation, which are described in detail in Angrist et al. (1996) and Stuart et al. (2008).

We are reasonably confident that the study meets four of the five assumptions. We are not convinced that we meet the exclusion restriction that states that there is no effect for students who do not fully participate. We believe, in short, that the noncomplier average causal effect (NACE) is not equal to zero. This is typical of evaluations of interventions with high cutoffs (Connell, 2009; Stuart et al., 2008). The most likely type of bias resulting from the exclusion restriction violation in this case is an overestimation of the CACE effect (Stuart et al., 2008). However, the effects of bias due to this violation can be mitigated by the inclusion of covariates that are predictive of participation (Jo, 2002).

We estimate the CACE with the *ivregress 2sls* command in Stata 15 (StataCorp, 2017). The first stage model predicts compliance (full participation) using the instrument (treatment assignment). The second stage predicts the outcome, given participation. The simultaneous estimation framework allows the user to calculate accurate standard errors that account for the uncertainty in the first stage model (Stuart et al., 2008). The benefit of the TSLS model is that it allows for the inclusion of baseline covariates that predict both participation and the outcome, which can help further reduce the amount of error in the estimation and possibly reduce bias due to exclusion restriction violations (Jo, 2002). We include the following covariates in both stages: age, race/ethnicity, gender, IEP status, ELL status, a baseline

measure of the outcome of interest, and a series of dummy variables representing the eight randomization blocks.

### *BASELINE EQUIVALENCE*

To examine baseline equivalence between the PGC-HS and control groups, we generate a model-based estimate of the standardized mean difference (SMD) between treatment and control groups on the pre-intervention covariates. Separate models are run, and estimates produced, for each of the variables selected for baseline equivalence for each outcome.[5] Where the baseline variable is continuous, the model is estimated with OLS and the standardized difference of means is calculated using the Hedges' *g* formula; where the baseline variable is dichotomous, the model is estimated with the LPM and the difference in the probability of the occurrence is calculated with the Cox Index formula.

### ITT AND INSTRUMENTAL VARIABLE SAMPLES

We were able to maintain low attrition (overall and differential) rates on all outcomes for our ITT sample. Across all outcome samples, the overall attrition rates for the ITT study ranged from 4.8 to 12.4% across outcomes and the differential attrition rates ranged from 0.0 to 2.3%. Similarly, in the instrumental variable study, which includes students with outcome data at six of the eight school blocks, overall attrition rates ranged from 4.7 to 12.5% and differential rates ranged from 0.1 to 2.9%.[6] These are well below the cautious boundary line for an acceptable threat of bias due to attrition, as outlined by the What Works Clearinghouse (U.S. Department of Education, 2020).[7] Because the attrition rates for both the ITT and instrumental variable studies were very low, we do not report baseline balance statistics for the treatment and control groups for these approaches.[8] Missing covariate data, including missing baseline data, were handled according to the techniques outlined by the National Center for Education Evaluation and Regional Assistance (May, Perez-Johnson, Haimson, Sattar, & Gleason, 2009). Because the overall and differential attrition was low for each reported outcome, we operate under the assumption that data are missing at random. As such, missing covariate data were treated using dummy variable adjustment according to guidance provided by Puma, Olsen, Bell, & Price (2009).[9]

### PRINCIPAL SCORE SAMPLES

Our benchmark approach to examining baseline equivalence between the PGC-HS and weighted control groups in the principal score analysis (described below) was to assess balance using complete case covariates; missing baseline data were not imputed. We follow the same procedure outlined above except that we add the principal score analytic weights to the model that estimates the standardized difference of means and proportions between treatment and control groups.



[5] Details of the model specifications used to estimate baseline balance statistics are included in the Supplementary Materials.
[6] Attrition rates for each outcome in the ITT and instrumental variable studies are included in the Supplemental Materials.
[7] We provide tables with the overall and differential attrition rates of the ITT and instrumental variable samples for each outcome in the Supplementary documentation available online.
[8] Per the What Works Clearinghouse Standards Handbook, Version. 4.1 guidance, we do not count the 181 students who were initially randomized at the ninth school block in our denominator when calculating attrition rates for the ITT sample. This is permissible because the event that affected the school (hurricane) and prevented the adequate implementation of the PGC-HS program affected students who were randomized to the treatment and control groups equally. However, to maintain transparency and provide additional justification for this decision, we did calculate attrition rates when these additional 181 students are included in the denominator. When these students are included in the denominator, overall attrition rates range from 16.1 to 22.7% and differential attrition rates range from 0.0 to 1.9%. Therefore, even if this loss of sample were to be included as overall attrition, differential attrition remains low enough that even by the conservative assumption standards, our ITT study samples remain within the acceptable attrition boundaries specified by the What Works Clearinghouse.
[9] Although our benchmark approach is to use dummy variable-adjusted baseline covariates to assess baseline equivalence and in our outcome models, we also tested the robustness of this approach by running sensitivity tests using complete case covariates. No substantive differences in equivalence statistics or impact estimates were observed.

# RESULTS

Baseline characteristics of the randomized study sample, PGC-HS compliers, and PGC-HS noncompliers are summarized in Table 1. Of the 680 ninth-grade students assigned to the PGC-HS (treatment) condition, 412 (61%) met the definition of compliance by attending at least 14 outreach sessions during their ninth-grade year. Among the PGC-HS group, compliers attended 18 outreach sessions, on average; noncompliers attended only 7. Compared with noncompliers, PGC-HS compliers were more likely to be White and attended approximately 10 more days of school in the eighth grade. Compliers were also more likely to report wanting (educational aspirations) or expecting to (educational expectations) obtain at least a four-year degree after high school at baseline.

## INTENT-TO-TREAT ANALYSIS

The results of the ITT analysis (Table 2) indicate that the opportunity to participate in PGC-HS did not have a statistically discernible effect on academic achievement and most noncognitive outcomes at the end of the ninth-grade year; however, impact estimates suggest that offering PGC-HS to ninth-grade students does have a positive effect on two behavioral and two noncognitive measures. Students assigned to the PGC-HS condition were five percent less likely to get suspended ($p$ = 0.020) or receive a disciplinary referral ($p$ = 0.069) compared with students assigned to the control condition, a difference that is statistically significant for suspensions and marginally significant on referrals. Within the noncognitive outcome domain, PGC-HS participants scored one tenth of a point higher (4.8 out of 7) on average on a measure of school attachment than students assigned to the control condition (4.7 out of 7; $p$ = 0.041). Additionally, compared with the control group, students in the PGC-HS group were five percent more likely to report they thought they would get at least a four-year degree after graduating high school ($p$ = 0.037). Otherwise, ITT results are statistically insignificant across the remaining outcomes.

[Table 2 goes here]

## PRINCIPAL SCORE ANALYSIS

Next, we estimate the CACE for PGC-HS with principal score sample weighting. Results, presented in Table 2, show that treatment students who participate fully (i.e., attend 14 sessions) in the PGC-HS program demonstrate significantly better academic achievement, behavioral, and noncognitive outcomes than those in the weighted comparison group. Specifically, PGC-HS compliers were seven percent less likely to be suspended ($p$ = 0.004) and six percent less likely to receive a disciplinary referral ($p$ = 0.036); they also achieved 0.15 point higher weighted GPAs ($p$ = 0.017), and scored between one tenth and two tenths of a point higher on several social and emotional measures, including growth mindset ($p$ = 0.046), decision-making skills ($p$ = 0.013), school attachment ($p$ = 0.001), and peer norms for academic achievement ($p$ = 0.001).[10] In addition, our analyses also suggest that PGC-HS compliers attended marginally more days of school ($p$ = 0.058), scored marginally higher on a measure of social competence ($p$ = 0.052), and were five percent more likely to report wanting to obtain a four-year degree ($p$ = 0.067) compared with a weighted control group.

The validity of the principal score analysis is predicated in part on the equivalence of the two contrasted groups. We can partially assess this by comparing the groups across an array of observed characteristics

[10] To assess how broadly we might interpret these results, we conducted an additional analysis that estimated the ATE effect of treatment assignment, using the analytic model described under the Intent-to-Treat analysis section, on all students at the six schools that implemented at least 14 outreach sessions and were subsequently included in CACE analyses. Results were substantively similar to the findings that were consistently observed by both CACE methods.

at baseline. In Table 3, we present baseline balance statistics, in the form of standardized differences of means, for the treatment and weighted comparison complier groups for a single outcome - number of days attended in ninth grade. We provide baseline balance statistics for only one outcome sample for brevity. Whereas each analytic sample is slightly different owing to differences in nonresponse to outcome questions or missing academic data, balance statistics presented in Table 3 are substantively the same for all outcome samples. Overall, baseline equivalence statistics show that the two groups are well balanced. The What Works Clearinghouse considers SMDs at or below 0.05 to indicate that the treatment and control samples are balanced. Where balance statistics are between 0.05 and 0.25, the What Works Clearinghouse recommends statistical adjustment by including the covariate in the outcome analytic model. Balance statistics at or above 0.25 indicate that balance was not achieved. As presented in Table 3, standardized differences between the two groups are at or below 0.05 for the majority of characteristics and are below 0.25 for all characteristics. As discussed in the Methods section, we include in our impact model all baseline demographic characteristics (age, race/ethnicity, gender), measures of disadvantage (ELL, IEP), and the specified baseline measure of the outcome, regardless of whether the SMD is below 0.05 or within the statistical adjustment range.

[Table 3 goes here]

### INSTRUMENTAL VARIABLE ANALYSIS

Results from the instrumental variable analyses are presented in Table 2. For the most part, the TSLS estimates corroborate principal score findings. Full participation in the PGC-HS program results in better academic achievement, behavioral, and noncognitive outcomes. Although the magnitude of the effect estimate is at times greater than that produced by the principal score analyses, the standard error also increases, moderating the statistical power of the analysis somewhat. Nevertheless, and consistent with principal score estimates, the instrumental variable CACE estimates indicate PGC-HS program participants are less likely to be suspended ($p$ = 0.013) or receive a disciplinary referral ($p$ = 0.041), achieve higher weighted GPAs ($p$ = 0.015), and score higher on several social and emotional measures, including decision-making skills ($p$ = 0.042), school attachment ($p$ = 0.009), and peer norms for academic achievement ($p$ = 0.011). Contrary to the principal score analyses, the instrumental variable CACE impact estimates do not support the hypothesis that participating fully in PGC-HS results in increased days in attendance at school ($p$ = 0.996), growth mindset score ($p$ = 0.277), or social competence score ($p$ = 0.659).

## DISCUSSION

Results from the ITT analysis demonstrate that the offer to participate in PGC-HS has some effect on measured outcomes (prosocial behavior, school attachment, and educational outlook), but that the effects are not prevalent across all outcome domains. This may be explained, at least in part, by poor implementation fidelity at two schools and individual variations in attendance at all schools. Findings from the two different CACE approaches (principal score and instrumental variable) provide evidence that when students do comply and receive a threshold level of the offered intervention, PGC-HS can improve academic achievement, prosocial behavior, and noncognitive outcomes more broadly. Indeed, sensitivity analyses that estimate the ATE among all participants who attended the six schools that offered the minimum threshold level of outreach sessions corroborate this assertion providing evidence

that the program can be effective, on average, when implemented well.[11] With regard to CACE analyses, we find that when ninth-grade students participate fully in the PGC-HS program, they are significantly less likely to get suspended or receive a disciplinary referral, achieve higher weighted GPAs, and score higher on measures of decision-making skills, school attachment, and peer norms for academic achievement than students in the control group. Additionally, findings from one CACE approach (principal score weighting) suggest the potential for additional promising impact on outcomes of attendance, growth mindset, educational aspirations, and social competence.

Both ITT and CACE findings indicate that the program is effective at curbing the likelihood that ninth-grade students will commit disciplinary infractions during their first year of high school. Specifically, ITT estimates indicate that students who were offered the PGC-HS program were suspended at lower rates than control students. Similarly, principal score-weighted regression results suggest an even greater effect. Only 12% of PGC-HS participants who received the full program were suspended during ninth grade, compared with 19% of control participants, a difference of 7%. ITT and CACE estimates also suggest that students in the PGC-HS condition were less likely to receive a disciplinary referral than control participants. In terms of academic achievement, principal score results show that PGC-HS compliers achieved a cumulative weighted GPA of 2.74 at the end of ninth grade, compared to a control group average of 2.59, a difference of 0.15 grade points. The two-stage instrumental variable model produced similar estimates of effect. These results are meaningful in their own right because both disciplinary infractions and GPA have been identified to be powerful early warning indicators of students who later drop out of high school (Bruce, Bridgeland, Fox, & Balfanz, 2011). As Bruce and colleagues explain, they represent two of the three primary predictor domains, sometimes referred to as the "ABCs" – attendance, behavior, and course performance. The findings that indicate students who are offered the program experience lower suspension and discipline referral rates on average, and that those who receive the full program achieve higher GPAs provide evidence that the program can be an effective intervention for helping ninth-grade students persist through and ultimately graduate from high school.

Additionally, in the ITT study, we found that students in the PGC-HS condition reported higher levels of attachment to their schools and were more likely to believe they would obtain a postsecondary degree than control students. What is more, the standardized magnitude (effect size) of these differences are moderately large for educational outcomes; specifically the difference in school attachment has an effect size of 0.09 and the effect size for educational expectations is 0.18.[12] The significant difference in measures of these affective and cognitive constructs indicate that regardless of dosage and fidelity levels, the program is having a robust positive effect on students' perception of belonging in school and their self-efficacy to persist, which are believed to be proximal (mediating) factors of academic success.

Whether these effects can meaningfully counteract the long term declines in engagement, efficacy, and motivation that begin in the ninth grade lies largely outside the scope of our study. However, we do observe declines in average school attachment scores from pre- to post-program for both treatment and control groups, which is consistent with the literature on ninth-grade shock. What is encouraging is that the treatment group's rate of decline is one third less than that of the control group and that we see

[11] We offer the ATE results at the six blocks that offered a minimum of 18 outreach sessions to their students in Table S6 in the Supplementary Materials. Findings from these sensitivity analyses indicate that when the program is implemented well (school offered at least 18 outreach sessions), students offered PGC-HS were less likely to get suspended or receive a disciplinary referral, achieved higher GPAs, and scored higher on measures of decision-making skills, school attachment, and peer norms for academic achievement than students who were offered the control condition. These findings were statistically significant at the 0.05 level.

[12] Effect sizes for the remaining ITT analyses are presented in Table S4 in the Supplementary Materials.

consistency in impact on school attachment across all of our analyses. Additionally, CACE findings consistently show that treatment compliers report practicing positive decision-making skills more frequently (e.g., *I stop and think about my options before making a decision*) and being a part of a social circle that shares positive academic norms (e.g., *The students I hang around with at school think that it's good to really like learning*) than control students who are weighted to look similar to compliers.

From a prevention policy and programming standpoint, these findings are of interest because they demonstrate that the program is effective for students who achieve a certain level of exposure. They also imply that the program could be more broadly impactful if implementation is improved such that all participating schools are able to offer the full curriculum (26 sessions) so that more students are given the opportunity to comply. Finally, they suggest that removing barriers to participation, and improving students' motivation to participate might result in broader impacts still. Taken together, findings support the basic hypotheses of the PGC-HS model that leveraging existing resources within schools (upper class mentors, faculty advisors) could be a low-cost and sustainable model that can help ninth-grade students persist through the difficult transition to high school and improve retention and graduation rates.

With a few and different assumptions, the CACE methods employed here estimate treatment effects for participants conditional on post-randomization selection without confounding self-selection bias. Although at least one of those assumptions is not convincingly met in the instrumental variable analysis, we have endeavored to mitigate the effects of violating this assumption. Further, because different assumptions underlie the two methods employed, and most of the findings (suspension, disciplinary referral, GPA, decision-making skills, school attachment, and peer norms) are essentially robust to those methods, we are reasonably confident in the effect estimates identified here.[13, 14]

This study contributes to the growing body of work regarding how and when researchers evaluating complex social and behavioral interventions can capitalize on the benefits of a randomized controlled trial to examine effects of varying dosage levels on outcomes, particularly when rich baseline data are available to predict compliance. By applying two different CACE approaches, each with a discrete set of statistical assumptions, and obtaining consistent results, we are able to more confidently infer that, as hypothesized, a high dosage level can lead to stronger program effects on academic, behavioral, and noncognitive outcomes.

## LIMITATIONS AND FUTURE WORK

As with any evaluation, the current investigation has several limitations that warrant mentioning. First, both the principal score and instrumental variable CACE methods employed here rest on untestable statistical assumptions. There is not much to be done about that, except that, following the advice of Jo

[13] Specifically, the principal score approach assumes principal ignorability and allows one to estimate the NACE, which informs how well the exclusion restriction holds. Conversely, the TSLS regression model allows one to examine how likely principal ignorability is to hold.

[14] To help make sense of the observed discrepancies between the two methods, we note two characteristics of our investigation that would suggest the principal score approach is the more valid method of estimating the causal effect of participating in the PGC-HS program. First, the threshold for compliance was set high at 16 or more outreach sessions. The exclusion restriction assumption in the instrumental variable approach asserts that any participation in the program below this threshold yields zero effect on outcomes. As noted, this assumption may be the hardest to justify and could have led to larger estimated point estimates of effect. Second, we collected extensive baseline data that measured pre-intervention outcomes and that we believe effectively predicted principal strata membership, in turn providing evidence that the principal ignorability assumption is satisfied. Given that previous research suggests that the propensity score approach performs best when the principal ignorability assumption is met and when the exclusion restriction assumption is not (Stuart & Jo, 2015), we would argue that the principal score approach is the more valid method in this particular case for calculating CACE estimates.

and Stuart (2009), we have used two different statistical techniques on the same data, to inform the validity of each other's underlying assumptions and assess the confidence in our results.

Another limitation is that program attendance data, which we rely upon to operationally define compliance, may be occasionally inaccurate. We do not believe this to be a systemic issue; however, implementation data suggest that there were a small number of outreach sessions offered at schools where attendance was not recorded or provided. Therefore, there may have been some PGC-HS participants who attended 14 or more sessions, who have been erroneously labeled as non-compliers. There are implications for both methods. In terms of the principal score approach, there are two concerns. First, incorrectly labeling treatment compliers as noncompliers decreases the analytic sample size (and therefore statistical power) because treatment noncompliers are assigned a weight of zero. Second, inadvertently grouping compliers into the noncomplier group could lead to inaccurate weighting of the comparison group since weights are based on one's probability of being a complier. Alternatively, in the two-stage regression model that aimed to estimate participation, given treatment assignment and subsequently the outcome, given participation, these mislabeled students could inflate the amount of error estimated in the models, making it harder to detect a significant difference in outcomes.

Finally, this study largely represents exploratory research in terms of its outcomes and analytic approaches. As part of the i3 grant process, two ITT confirmatory contrasts (ninth-grade attendance and credits earned) were prespecified in detail by the research team in the i3 Evaluation Design Summary, including the research questions, outcome operationalization, assignment procedures, data collection methods, and analytic approach. Beyond attendance and credits, the remaining outcome measures were noted as exploratory because they were either considered alternative indicators of persistence (promotion, GPA, discipline) or theoretical antecedents of behavior (noncognitive outcomes). As a result, outcome operationalization was not prespecified in detail. Similarly, the CACE analyses were not prespecified and were endeavored as a means of providing additional context for the ITT findings and provide evidence of effectiveness when study participants receive the full scope of programming.

This study aimed to estimate the CACE of participating fully in the PGC-HS program using two different approaches, one based on propensity score methods and one based on instrumental variable TSLS regression. A third approach that is becoming more commonly used, ASPES, was not utilized for this study, but future work using this method could provide additional clarity and evidence of effectiveness if results were found to be consistent. In addition, this study examined the effect of participation on outcomes measured at the end of ninth grade soon after the intervention ended. Future work that examines the effect of PGC-HS participation on long-term outcomes measured at the end of subsequent academic years could provide additional evidence for the program's effect on retention in and graduation from high school.

# REFERENCES

Allensworth, E., & Easton, J. (2005). *The on-track indicator as a predictor of high school graduation.* Chicago, IL: University of Chicago Consortium on School Research.

Alliance for Excellent Education. (2013). *Saving futures, saving dollars: The impact of education on crime reduction and earnings.* Retrieved from https://all4ed.org/wpcontent/uploads/2013/09/SavingFutures.pdf

Angrist, J., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, *90*(430), 431–442. doi:10.2307/2291054

Angrist, J., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–472. doi:10.2307/2291629

Appleton, J., Christenson, S., & Furlong, M. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, *45*, 369–386. doi:10.1002/pits.20303

Bell, S. H., & Peck, L. R. (2016). On the "how" of social experiments: Experimental designs for getting inside the black box. *New Directions for Evaluation*, *2016*(152), 97–107. doi:10.1002/ev.20210

Bernat, D. H., & Resnick, M. D. (2009). Connectedness in the lives of adolescents. In R. J. DiClemente, J. S. Santelli, & R. A. Crosby (Eds.), *Adolescent health: Understanding and preventing risk behaviors* (pp. 375–389). Hoboken, NJ: Jossey-Bass/Wiley.

Black, M. M., Bentley, M. E., Papas, M. A., Oberlander, S., Teti, L. O., McNary, S., . . . O'Connell, M. (2006). Delaying second births among adolescent mothers: A randomized, controlled trial of a homebased mentoring program. *Pediatrics*, *118*(4), e1087–e1099. doi:10.1542/peds.2005-2318

Bruce, M., Bridgeland, J. M., Fox, J. H., & Balfanz, R. 2011. *On track for success: The use of early warning indicator and intervention systems to build a grad nation*. Washington, DC: Civic Enterprises.

Centers for Disease Control and Prevention. 2009. *School connectedness: Strategies for increasing protective factors among youth*. Atlanta, GA: U.S. Department of Health and Human Services.

Chan, W. Y., Kuperminc, G. P., Seitz, S., Wilson, C., & Khatib, N. (2020). School-based group mentoring and academic outcomes in vulnerable high-school students. *Youth & Society*, *52*(7), 1220–1237. doi:10.1177/0044118X19864834

Christenson, S. L., Reschly, A. L., & Wylie, C. (Eds.). (2012). *Handbook of research on student engagement*. doi.org/10.1007/978-1-4614-2018-7

Cohen, J. S., & Smerdon, B. A. (2009) Tightening the dropout tourniquet: Easing the transition from middle to high school. *Preventing School Failure: Alternative Education for Children and Youth*, *53*(3), 177–184. doi:10.3200/PSFL.53.3.177-184

Connell, A. M. (2009). Employing complier average causal effect analytic methods to examine effects of randomized encouragement trials. *The American Journal of Drug and Alcohol Abuse*, *35*(4), 253–259. doi:10.1080/00952990903005882

Dunn, G., Maracy, M., Dowrick, C., Ayuso-Mateos, J. L., Dalgard, O. S., Page, H., . . . Wilkinson, G. (2003). Estimating psychological treatment effects from a randomised controlled trial with both noncompliance and loss to follow-up. *The British Journal of Psychiatry*, *183*, 323–331. doi:10.1192/bjp.183.4.323

Easton, J. Q., Johnson, E., & Sartain, L. (2017). *The predictive power of ninth-grade GPA.* Chicago, IL: University of Chicago Consortium on School Research.

Follmann, D. A. (2000) On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association*, *95*(452), 1101–1109.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29. doi:10.1111/j.0006-341x.2002.00021.x

Fredricks, J., Blumenfeld, P., & Paris, A. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109.

Hill, J., Waldfogel, J., & Brooks-Gunn, J. (2002). Differential effects of high-quality child care. *Journal of Policy Analysis and Management*, *21*(4), 601–627. doi:10.1002/pam.10077

Hill, J., Waldfogel, J., & Brooks-Gunn, J. (2003) Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology*, *39*(4), 730–744.

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., . . . Dilig, R. (2020). *The Condition of Education 2020* (NCES 2020-144). Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo. asp?pubid=2020144

Jo, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods*, *7*, 178–193.

Jo, B., & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, *28*, 2857–2875. doi:10.1002/sim.3669

Keaton, P. (2012). *Public elementary and secondary school student enrollment and staff counts from the common core of data: School year 2010–11* (NCES 2012-327). Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from http://nces.ed.gov/pubsearch

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Moulton, S., Peck, L., & Greeney, A. (2018). Analyzing the influence of dosage in social experiments with application to the supporting healthy marriage program. *American Journal of Evaluation*, *39*(2), 257–277. doi:10.1177/1098214017698566

Peck, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation*, *24*(2), 157–187. doi:10.1016/S1098-2140(03)00031-6

Peck, L. R. (Ed.). (2016). Social experiments in practice: The what, why, when, where, and how of experimental design and analysis [Special issue]. *New Directions for Evaluation*, *2016*(152).

Pharris-Ciurej, N., Hirschman, C., & Willhoft, J. (2012). The 9th grade shock and the high school dropout crisis. *Social Science Research*, *41*(3), 709–730. doi.org/10.1016/j.ssresearch.2011.11.014

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Rallis, S. (Ed.). (2015). Unpacking the "black box" of social programs and policies [Special issue]. *American Journal of Evaluation*, *36*(4).

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Schochet, P. Z., & Chiang, H. S. (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics*, *36*(3), 307–345.

Smith, J. S. (2006). *Research summary: Transition from middle school to high school*. Retrieved from http://images.pcmac.org/Uploads/Triton/Triton/Divisions/DocumentsCategories/Documents/Transition_from_MStoHS-1.pdf

StataCorp. 2015. Stata Statistical Software (Release 14). College Station, TX: StataCorp LP.

StataCorp. 2017. Stata Statistical Software (Release 15). College Station, TX: StataCorp LLC.

Stuart, E. A., & Jo, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Statistical Methods in Medical Research*, *24*(6), 657–674. doi:10.1177/0962280211421840

Stuart, E. A., Perry, D. F., Huynh-Nhu, L., & Ialongo, N. S. (2008). Estimating intervention effects of prevention programs: Accounting for noncompliance. *Prevention Science*, *9*, 288–298. doi:10.1007/s11121-008-0104-y

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. (March 2014). *Dropout prevention evidence review protocol, Version 3*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_dp_protocol_v3.0.pdf

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. (January 2020). *Standards Handbook, Version 4.1*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

U.S. Department of Education, Office of Innovation and Improvement, Applications for New Awards; Investing in Innovation Fund – Development Grants, (March 2015). Retrieved from https://www.govinfo.gov/content/pkg/FR-2015-03-30/pdf/2015-07213.pdf

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service (2017). *Issue Brief: Mentoring*. https://www2.ed.gov/rschstat/eval/high-school/mentoring.pdf

Zimmerman, M. A., Bingenheimer, J. B., & Notaro, P. C. (2002) Natural mentors and adolescent resiliency: A study with urban youth. (2002). *American Journal of Community Psychology*, *30*, 221–243. doi:10.1023/A:1014632911622

# TABLES

*Table 1. Baseline Characteristics by Intervention Group and Compliance Status*

| | Randomized Sample | | | | PGC-HS Group | | | |
|---|---|---|---|---|---|---|---|---|
| | PGC-HS (*n* = 680) | | Control (*n* = 671) | | Non-Compliers (*n* = 268) | | Compliers (*n* = 412) | |
| **Baseline Characteristic** | **M** | **SD** | **M** | **SD** | **M** | **SD** | **M** | **SD** |
| Age at baseline | 14.59 | 0.50 | 14.61 | 0.50 | 14.66 | 0.53 | 14.55 | 0.47 |
| Female | 47.94% | 0.50 | 45.75% | 0.50 | 45.90% | 0.50 | 49.27% | 0.50 |
| White | 42.79% | 0.50 | 42.77% | 0.50 | 29.10% | 0.46 | 51.70% | 0.50 |
| Black | 36.03% | 0.48 | 36.21% | 0.48 | 41.42% | 0.49 | 32.52% | 0.47 |
| Other | 18.97% | 0.39 | 19.08% | 0.39 | 24.25% | 0.43 | 15.53% | 0.36 |
| Hispanic | 12.50% | 0.33 | 11.48% | 0.32 | 12.31% | 0.33 | 12.62% | 0.33 |
| ELL student | 1.62% | 0.13 | 1.94% | 0.14 | 1.49% | 0.12 | 1.70% | 0.13 |
| IEP student | 12.65% | 0.33 | 11.03% | 0.31 | 12.69% | 0.33 | 12.62% | 0.33 |
| Number of days attended in 8th grade | 162.55 | 20.80 | 160.46 | 23.18 | 156.79 | 24.14 | 166.31 | 17.34 |
| Number of classes passed in 8th grade | 7.19 | 1.09 | 7.17 | 1.09 | 7.19 | 1.21 | 7.18 | 1.01 |
| **Noncognitive Outcomes** | | | | | | | | |
| Self-efficacy in goal setting | 5.76 | 0.94 | 5.72 | 1.02 | 5.64 | 0.98 | 5.84 | 0.90 |
| Growth mindset | 5.46 | 0.81 | 5.47 | 0.80 | 5.41 | 0.83 | 5.49 | 0.79 |
| Grit | 5.19 | 0.98 | 5.20 | 0.97 | 5.11 | 0.94 | 5.24 | 1.00 |
| Decision-making skills | 5.18 | 1.25 | 5.13 | 1.31 | 5.12 | 1.28 | 5.22 | 1.24 |
| Educational ambitions | 6.34 | 0.80 | 6.36 | 0.82 | 6.31 | 0.85 | 6.36 | 0.77 |
| Peer connection | 5.82 | 1.25 | 5.84 | 1.27 | 5.66 | 1.34 | 5.93 | 1.17 |
| School attachment | 4.99 | 1.05 | 5.00 | 1.10 | 4.82 | 1.05 | 5.10 | 1.03 |
| Social competence | 5.51 | 1.14 | 5.52 | 1.15 | 5.30 | 1.13 | 5.65 | 1.13 |
| Peer norms for academic achievement | 4.84 | 0.95 | 4.79 | 0.97 | 4.71 | 0.90 | 4.93 | 0.98 |
| Educational aspirations | 72.06% | 0.45 | 70.94% | 0.45 | 64.55% | 0.48 | 76.94% | 0.42 |
| Educational expectations | 56.18% | 0.50 | 56.78% | 0.50 | 45.90% | 0.50 | 62.86% | 0.48 |
| **School** | | | | | | | | |
| Block 1 | 14.41% | 0.35 | 14.90% | 0.36 | 6.72% | 0.25 | 19.42% | 0.40 |
| Block 2 | 5.88% | 0.24 | 5.81% | 0.23 | 6.72% | 0.25 | 5.34% | 0.23 |
| Block 3 | 14.41% | 0.35 | 14.46% | 0.35 | 10.45% | 0.31 | 16.99% | 0.38 |
| Block 4 | 16.03% | 0.37 | 15.80% | 0.36 | 13.43% | 0.34 | 17.72% | 0.38 |
| Block 5 | 15.29% | 0.36 | 15.05% | 0.36 | 4.85% | 0.22 | 22.09% | 0.42 |
| Block 6 | 10.00% | 0.30 | 10.13% | 0.30 | 25.37% | 0.44 | 0.00% | 0.00 |
| Block 7 | 13.53% | 0.34 | 13.71% | 0.34 | 5.97% | 0.24 | 18.45% | 0.39 |
| Block 8 | 10.44% | 0.31 | 10.13% | 0.30 | 26.49% | 0.44 | 0.00% | 0.00 |

*Note: PGC-HS group non-compliers are members of the treatment group who attended fewer than 14 outreach sessions; PGC-HS group compliers are members of the treatment group who attended 14 or more outreach sessions.*

*Table 2. Results of ITT and Complier Analyses*

| | | | Complier Effects | | | |
|---|---|---|---|---|---|---|
| | ITT Impact | | Principal Score | | Instrumental Variable | |
| **Outcome** | **N** | **β (SE)** | **N** | **β (SE)** | **N** | **β (SE)** |
| Days attended | 1,255 | 0.306(0.676) | 828 | 1.415(0.744)~ | 1,004 | –0.004(0.893) |
| Was suspended | 1,213 | –0.049(0.021)* | 808 | –0.073(0.025)** | 963 | –0.070(0.028)* |
| Received disciplinary referral | 1,213 | –0.047(0.026)~ | 808 | –0.065(0.031)* | 963 | –0.071(0.035)* |
| Received detention | 1,219 | –0.012(0.021) | 813 | –0.012(0.026) | 969 | –0.022(0.029) |
| Credits earned | 1,275 | –0.019(0.073) | 838 | 0.061(0.080) | 1,019 | 0.056(0.106) |
| Promoted to 10th grade | 1,286 | –0.008(0.016) | 842 | 0.010(0.018) | 1,025 | 0.007(0.022) |
| Weighted GPA | 1,280 | 0.071(0.050) | 843 | 0.146(0.061)* | 1,025 | 0.173(0.071)* |
| Self-efficacy in goal setting | 1,212 | 0.011(0.049) | 794 | 0.006(0.057) | 963 | 0.016(0.069) |
| Growth mindset | 1,206 | 0.009(0.043) | 786 | 0.099(0.050)* | 955 | 0.066(0.061) |
| Grit | 1,199 | –0.009(0.050) | 773 | 0.021(0.061) | 950 | 0.053(0.071) |
| Decision-making skills | 1,220 | 0.052(0.067) | 804 | 0.195(0.078)* | 970 | 0.183(0.090)* |
| Educational ambitions | 1,219 | –0.043(0.044) | 797 | 0.039(0.054) | 967 | 0.009(0.062) |
| Peer connection | 1,227 | 0.028(0.065) | 798 | 0.114(0.078) | 973 | 0.058(0.089) |
| School attachment | 1,189 | 0.111(0.054)* | 786 | 0.229(0.067)** | 947 | 0.198(0.075)** |
| Social competence | 1,222 | 0.013(0.059) | 802 | 0.133(0.068)~ | 969 | 0.035(0.079) |
| Peer norms for academic achievement | 1,184 | 0.073(0.049) | 742 | 0.195(0.060)** | 942 | 0.175(0.069)* |
| Educational aspirations | 1,210 | 0.026(0.021) | 792 | 0.047(0.026)~ | 965 | 0.026(0.029) |
| Educational expectations | 1,213 | 0.049(0.023)* | 790 | 0.041(0.028) | 965 | 0.040(0.033) |

*Note: ~p < 0.1, *p < 0.05, **p < 0.01.*

*Table 3. Baseline Characteristics and Balance Statistics for Principal Score Days Attended Analytic Sample*

| Baseline Characteristic | PGC-HS Compliers (*n* = 397) | Control Compliers (*n* = 431) | Standardized Mean Difference |
|---|---|---|---|
| Age at baseline | 14.55 | 14.57 | –0.04 |
| Female | 49.12% | 50.04% | –0.03 |
| White | 52.90% | 51.50% | 0.07 |
| Black | 31.49% | 30.74% | –0.01 |
| Other | 15.62% | 17.75% | –0.09 |
| Hispanic | 12.59% | 13.02% | –0.01 |
| ELL student | 1.76% | 2.12% | –0.11 |
| IEP student | 12.09% | 10.59% | 0.09 |
| Number of days attended in 8$^{th}$ grade | 166.59 | 165.27 | 0.07 |
| Number of classes passed in 8$^{th}$ grade | 7.19 | 7.15 | 0.02 |
| Self-efficacy in goal setting | 5.87 | 5.73 | 0.14 |
| Growth mindset | 5.51 | 5.48 | 0.04 |
| Grit | 5.25 | 5.20 | 0.05 |
| Decision-making skills | 5.24 | 5.20 | 0.04 |
| Educational ambitions | 6.37 | 6.32 | 0.05 |
| Peer connection | 5.95 | 5.93 | 0.02 |
| School attachment | 5.12 | 5.09 | 0.03 |
| Social competence | 5.66 | 5.60 | 0.05 |
| Peer norms for academic achievement | 4.93 | 4.82 | 0.10 |
| Educational aspirations | 79.23% | 79.02% | 0.00 |
| Educational expectations | 65.21% | 62.66% | 0.06 |
| Block 1 | 19.90% | 20.29% | –0.01 |
| Block 2 | 5.04% | 5.13% | –0.01 |
| Block 3 | 17.38% | 18.71% | –0.05 |
| Block 4 | 16.88% | 17.68% | –0.03 |
| Block 5 | 22.17% | 21.19% | 0.03 |
| Block 7 | 18.64% | 17.00% | 0.07 |

*Note: The following covariates are included in the benchmark model that assesses each outcome: age at baseline, gender, race/ethnicity, ELL and IEP student indicators, and randomization blocks. Number of days attended in 8$^{th}$ grade is used as a baseline measure of the outcome (or proxy) for the outcomes days attended in 9$^{th}$ grade, suspension, detention, and disciplinary referral. Number of classes passed in 8$^{th}$ grade is used as a baseline measure of the outcome (or proxy) for the outcomes credits earned in 9$^{th}$ grade, weighted GPA, and promotion to 10$^{th}$ grade. Baseline measures of each noncognitive outcome are included in the analytic models that assess each outcome at the end of 9$^{th}$ grade. A detailed explanation of methods used to calculate SMD are included in the Supplemental Materials available online.*

# SUPPLEMENTAL ONLINE MATERIALS

## TECHNICAL DETAILS

### *ASSIGNMENT PROCEDURES*

The impact study was a randomized controlled trial (RCT) where the unit of randomization and the unit of analysis were the individual student. For each study school, if a participant met all five eligibility criteria, we assigned them a unique study ID number and a random allocation to either the treatment or control condition at a 1:1 ratio where each student had a 50% chance of being assigned to either condition. Blocking was done at the school level. For schools that participated in the study for two years, separate blocks were created for each school year. Randomization lists were created for each study school and were constructed using the random allocation (*ralloc*) command in Stata 14. The command generates a sequence of treatment/control string codes in alternating blocks ranging from 2 to 10 cases. An analyst sorted each school's list of eligible students alphabetically by last then first name, then copied and pasted the randomization sequence to the list of eligible participants. Students were considered enrolled in the intent to treat (ITT) sample at the point of random assignment into either the treatment or control condition.

Once each eligible student at a school was assigned to either the treatment or control condition (described above), an analyst sent a copy to the school's Stakeholder Team Coordinator (STC) with instructions for scheduling students into PGC-HS outreach sessions according to their assignment. The school's STC was responsible for placing treatment participants into the Peer Group Connection – High School (PGC-HS) program and forming the peer groups. Once peer groups were constructed, the STC sent staff at Center for Supportive Schools (CSS) and the evaluation team the final roster of PGC-HS participants via a shared Google Drive.[15] Evaluators monitored fidelity to random assignment using these program rosters. Evaluators and CSS also monitored program attendance using the CSS-created *Attendance Tracker* to assess whether treatment students were attending the program as planned and whether any control students were placed in the program by the school.

## ATTRITION

In this section, we provide tables detailing the overall and differential attrition rates for each of the analytic samples used in the ITT (Table S1) and instrumental variable (Table S3) analyses. In each table, we provide the number of PGC-HS and non-PGC-HS participants who were randomized and in the analytic samples, followed by the overall attrition, or pooled sample loss, and the differential attrition, or the difference in sample loss between the PGC-HS and non-PGC-HS groups. Table S2 details the calculations when the ninth school block, which was excluded from the RCT and our benchmark ITT results, is included in the denominator of the ITT calculations.



[15] Because school STCs often had multiple responsibilities at the beginning of program implementation, participant lists were infrequently sent to CSS and the evaluation team until the program had been underway for a month or more.

*Table S1. Randomized and Analytic Samples of ITT Analyses*

| | Number Randomized | | Analytic Sample | | | |
|---|---|---|---|---|---|---|
| | PGC | Non-PGC | PGC | Non-PGC | Overall Attrition | Differential Attrition |
| Days attended | 680 | 671 | 627 | 628 | 7.1% | 1.4% |
| Was suspended | 680 | 671 | 609 | 604 | 10.2% | 0.5% |
| Received disciplinary referral | 680 | 671 | 609 | 604 | 10.2% | 0.5% |
| Received detention | 680 | 671 | 612 | 607 | 9.8% | 0.5% |
| Credits earned | 680 | 671 | 634 | 641 | 5.6% | 2.3% |
| Promoted to 10th grade | 680 | 671 | 642 | 644 | 4.8% | 1.6% |
| Weighted GPA | 680 | 671 | 639 | 641 | 5.3% | 1.6% |
| Self-efficacy in goal setting | 680 | 671 | 612 | 600 | 10.3% | –0.6% |
| Growth mindset | 680 | 671 | 611 | 595 | 10.7% | –1.2% |
| Grit | 680 | 671 | 610 | 589 | 11.3% | –1.9% |
| Decision-making skills | 680 | 671 | 613 | 607 | 9.7% | 0.3% |
| Educational ambitions | 680 | 671 | 615 | 604 | 9.8% | –0.4% |
| Peer connection | 680 | 671 | 620 | 607 | 9.2% | –0.7% |
| School attachment | 680 | 671 | 600 | 589 | 12.0% | –0.5% |
| Social competence | 680 | 671 | 618 | 604 | 9.5% | –0.9% |
| Peer norms for academic achievement | 680 | 671 | 593 | 591 | 12.4% | 0.9% |
| Educational aspirations | 680 | 671 | 609 | 601 | 10.4% | 0.0% |
| Educational expectations | 680 | 671 | 615 | 598 | 10.2% | –1.3% |

*Table S2. Randomized and Analytic Samples of ITT Analyses with Ninth Block Counted Toward Attrition*[16]

| | Number Randomized | | Analytic Sample | | | |
|---|---|---|---|---|---|---|
| | PGC | Non-PGC | PGC | Non-PGC | Overall Attrition | Differential Attrition |
| Days attended | 770 | 762 | 627 | 628 | 18.1% | 1.0% |
| Was suspended | 770 | 762 | 609 | 604 | 20.8% | 0.2% |
| Received disciplinary referral | 770 | 762 | 609 | 604 | 20.8% | 0.2% |
| Received detention | 770 | 762 | 612 | 607 | 20.4% | 0.2% |
| Credits earned | 770 | 762 | 634 | 641 | 16.8% | 1.8% |
| Promoted to 10th grade | 770 | 762 | 642 | 644 | 16.1% | 1.1% |
| Weighted GPA | 770 | 762 | 639 | 641 | 16.4% | 1.1% |
| Self-efficacy in goal setting | 770 | 762 | 612 | 600 | 20.9% | –0.7% |
| Growth mindset | 770 | 762 | 611 | 595 | 21.3% | –1.3% |
| Grit | 770 | 762 | 610 | 589 | 21.7% | –1.9% |
| Decision-making skills | 770 | 762 | 613 | 607 | 20.4% | 0.0% |
| Educational ambitions | 770 | 762 | 615 | 604 | 20.4% | –0.6% |
| Peer connection | 770 | 762 | 620 | 607 | 19.9% | –0.9% |
| School attachment | 770 | 762 | 600 | 589 | 22.4% | –0.6% |
| Social competence | 770 | 762 | 618 | 604 | 20.2% | –1.0% |
| Peer norms for academic achievement | 770 | 762 | 593 | 591 | 22.7% | 0.5% |
| Educational aspirations | 770 | 762 | 609 | 601 | 21.0% | –0.2% |
| Educational expectations | 770 | 762 | 615 | 598 | 20.8% | –1.4% |



[16]As described in the Method section of the paper, we do not consider these attrition statistics to be the benchmark calculations – those are presented in Table S1 – but we include them here for to verify the claim that our ITT analytic samples meet the low-attrition designation regardless of whether the loss of the ninth school sample is considered attrition or not.

*Table S3. Random and Analytic Samples of Instrumental Variable Analyses*

| | Number Randomized | | Analytic Sample | | | |
|---|---|---|---|---|---|---|
| | PGC | Non-PGC | PGC | Non-PGC | Overall Attrition | Differential Attrition |
| Days attended | 541 | 535 | 497 | 507 | 6.7% | 2.9% |
| Was suspended | 541 | 535 | 480 | 483 | 10.5% | 1.6% |
| Received disciplinary referral | 541 | 535 | 480 | 483 | 10.5% | 1.6% |
| Received detention | 541 | 535 | 483 | 486 | 9.9% | 1.6% |
| Credits earned | 541 | 535 | 502 | 517 | 5.3% | 3.8% |
| Promoted to 10th grade | 541 | 535 | 508 | 517 | 4.7% | 2.7% |
| Weighted GPA | 541 | 535 | 508 | 517 | 4.7% | 2.7% |
| Self-efficacy in goal setting | 541 | 535 | 485 | 478 | 10.5% | –0.3% |
| Growth mindset | 541 | 535 | 482 | 473 | 11.2% | –0.7% |
| Grit | 541 | 535 | 480 | 470 | 11.7% | –0.9% |
| Decision-making skills | 541 | 535 | 484 | 486 | 9.9% | 1.4% |
| Educational ambitions | 541 | 535 | 486 | 481 | 10.1% | 0.1% |
| Peer connection | 541 | 535 | 489 | 484 | 9.6% | 0.1% |
| School attachment | 541 | 535 | 475 | 472 | 12.0% | 0.4% |
| Social competence | 541 | 535 | 487 | 482 | 9.9% | 0.1% |
| Peer norms for academic achievement | 541 | 535 | 469 | 473 | 12.5% | 1.7% |
| Educational aspirations | 541 | 535 | 484 | 481 | 10.3% | 0.4% |
| Educational expectations | 541 | 535 | 486 | 479 | 10.3% | –0.3% |

*Note: The instrumental variable analysis included all participants with outcome data in six of the eight randomization blocks.*
*Note that the above calculations also reflect attrition rates for the ATE analyses conducted with high fidelity sites (see Table S6 below).*

### ANALYSIS WITH MISSING DATA

We did not impute any missing outcome data. Impact analysis samples include only those observations that have non-missing outcome (post-intervention) data. As a result, the analytic sample for each research question varies slightly. Missing covariate data, including missing baseline data, were handled according to the techniques outlined by the National Center for Education Evaluation and Regional Assistance (May et al., 2009). Because the overall and differential attrition was low for each reported outcome (both primary and exploratory), we operate under the assumption that data are missing at random. As such, missing covariate data, including baseline outcome data, were treated using dummy variable adjustment according to guidance provided by Puma et al. (2009).

## EFFECT SIZE AND POOLED STANDARD DEVIATION

Table S4 presents the detailed analytic results of the ITT analyses. For each outcome assessed, we provide the number of participants included in the analytic sample and the group mean of the outcome, by treatment group, the impact coefficient and its standard error (se), as well as the *p*-value and calculated effect size of the impact estimate.

*Table S4. Detailed ITT Impact Results*

| | PGC | | Control | | | | |
|---|---|---|---|---|---|---|---|
| **Outcome** | **N** | **Mean (SD)** | **N** | **Mean (SD)** | **β (se)** | ***p*-value** | **Effect Size** |
| Days attended | 627 | 166.33(12.00) | 628 | 165.73(13.63) | 0.31(0.68) | 0.651 | 0.02 |
| Was suspended | 609 | 0.14(0.35) | 604 | 0.20(0.40) | –0.05(0.02) | 0.020 | –0.24 |
| Received disciplinary referral | 609 | 0.34(0.48) | 604 | 0.40(0.49) | –0.05(0.03) | 0.069 | –0.14 |
| Received detention | 612 | 0.18(0.39) | 607 | 0.20(0.40) | –0.01(0.02) | 0.576 | –0.04 |
| Credits earned | 634 | 7.28(1.44) | 641 | 7.27(1.47) | –0.02(0.07) | 0.790 | –0.01 |
| Promoted to 10th grade | 642 | 0.88(0.32) | 644 | 0.89(0.31) | –0.01(0.02) | 0.644 | –0.03 |
| Weighted GPA | 639 | 2.71(1.07) | 641 | 2.63(1.03) | 0.07(0.05) | 0.155 | 0.07 |
| Self-efficacy in goal setting | 612 | 5.81(1.02) | 600 | 5.74(1.09) | 0.01(0.05) | 0.827 | 0.01 |
| Growth mindset | 611 | 5.41(0.91) | 595 | 5.41(0.85) | 0.01(0.04) | 0.844 | 0.01 |
| Grit | 610 | 5.25(1.01) | 589 | 5.26(1.00) | –0.01(0.05) | 0.861 | –0.01 |
| Decision-making skills | 613 | 5.17(1.40) | 607 | 5.05(1.37) | 0.05(0.07) | 0.438 | 0.04 |
| Educational ambitions | 615 | 6.21(0.98) | 604 | 6.25(0.93) | –0.04(0.04) | 0.326 | –0.04 |
| Peer connection | 620 | 5.80(1.32) | 607 | 5.79(1.35) | 0.03(0.07) | 0.674 | 0.02 |
| School attachment | 600 | 4.83(1.14) | 589 | 4.72(1.20) | 0.11(0.05) | 0.041 | 0.09 |
| Social competence | 618 | 5.45(1.26) | 604 | 5.44(1.23) | 0.01(0.06) | 0.832 | 0.01 |
| Peer norms for academic achievement | 593 | 4.76(1.07) | 591 | 4.65(1.02) | 0.07(0.05) | 0.140 | 0.07 |
| Educational aspirations | 609 | 0.76(0.43) | 601 | 0.73(0.44) | 0.03(0.02) | 0.213 | 0.13 |
| Educational expectations | 615 | 0.61(0.49) | 598 | 0.56(0.50) | 0.05(0.02) | 0.037 | 0.18 |

## QUESTIONNAIRE SCALE RELIABILITY

Table S5 presents the reliability statistic (Cronbach's alpha) for each of the scaled outcomes at baseline.

*Table S5. Baseline Noncognitive Scale Reliability*

| | **N** | **Mean** | **SD** | **Cronbach's Alpha** |
|---|---|---|---|---|
| Self-efficacy in goal setting | 1,351 | 5.74 | 0.98 | 0.84 |
| Growth mindset | 1,351 | 5.46 | 0.80 | 0.63 |
| Grit | 1,351 | 5.20 | 0.97 | 0.73 |
| Decision-making skills | 1,351 | 5.16 | 1.28 | 0.83 |
| Educational ambitions | 1,351 | 6.35 | 0.81 | 0.84 |
| Peer connection | 1,351 | 5.83 | 1.26 | 0.90 |
| School attachment | 1,351 | 5.00 | 1.08 | 0.81 |
| Social competence | 1,351 | 5.51 | 1.15 | 0.81 |
| Peer norms for academic achievement | 1,351 | 4.81 | 0.96 | 0.80 |

## ATE RESULTS – SCHOOLS THAT MET FIDELITY THRESHOLD

Below we provide the results of analyses that estimated the average treatment effect (ATE) on all students who attended school at the six locations that implemented PGC such that they met the minimum threshold for fidelity by offering at least 18 outreach sessions to students. Of the eight school blocks that implemented PGC-HS, two did not implement the program to the minimum recommended threshold level. By presenting these results, we aim to offer additional context for PGC-HS' potential for impact when implemented well, beyond the ITT impact estimates on the full randomized sample. Note that the attrition calculations for each outcome are the same as those presented in Table S3 above.

*Table S6. Detailed Average Program Impact at High Fidelity Implementation Sites*

| | PGC | | Control | | | | |
|---|---|---|---|---|---|---|---|
| **Outcome** | **N** | **Mean (SD)** | **N** | **Mean (SD)** | **β (se)** | ***p*-value** | **Effect Size** |
| Days attended | 497 | 167.79(11.82) | 507 | 167.48(12.12) | 0.00(0.76) | 0.996 | 0.00 |
| Was suspended | 480 | 0.13(0.34) | 483 | 0.19(0.39) | –0.06(0.02) | 0.014 | –0.29 |
| Received disciplinary referral | 480 | 0.34(0.47) | 483 | 0.40(0.49) | –0.06(0.03) | 0.043 | –0.19 |
| Received detention | 483 | 0.19(0.39) | 486 | 0.21(0.41) | –0.02(0.02) | 0.452 | –0.07 |
| Credits earned | 502 | 7.30(1.44) | 517 | 7.24(1.53) | 0.04(0.08) | 0.601 | 0.03 |
| Promoted to 10$^{th}$ grade | 508 | 0.91(0.29) | 517 | 0.90(0.30) | 0.01(0.02) | 0.748 | 0.10 |
| Weighted GPA | 508 | 2.81(1.08) | 517 | 2.69(1.06) | 0.13(0.06) | 0.017 | 0.13 |
| Self-efficacy in goal setting | 485 | 5.82(0.98) | 478 | 5.72(1.08) | 0.01(0.05) | 0.815 | 0.01 |
| Growth mindset | 482 | 5.48(0.87) | 473 | 5.42(0.84) | 0.05(0.05) | 0.275 | 0.06 |
| Grit | 480 | 5.29(1.00) | 470 | 5.22(1.01) | 0.04(0.06) | 0.455 | 0.04 |
| Decision-making skills | 484 | 5.25(1.33) | 486 | 5.04(1.35) | 0.14(0.07) | 0.044 | 0.11 |
| Educational ambitions | 486 | 6.22(0.96) | 481 | 6.20(0.96) | 0.01(0.05) | 0.884 | 0.01 |
| Peer connection | 489 | 5.87(1.30) | 484 | 5.83(1.29) | 0.05(0.07) | 0.521 | 0.04 |
| School attachment | 475 | 4.96(1.13) | 472 | 4.77(1.20) | 0.15(0.06) | 0.010 | 0.13 |
| Social competence | 487 | 5.60 (1.14) | 482 | 5.53(1.19) | 0.03(0.06) | 0.662 | 0.02 |
| Peer norms for academic achievement | 469 | 4.85 (1.06) | 473 | 4.65(1.03) | 0.14(0.06) | 0.012 | 0.13 |
| Educational aspirations | 484 | 0.76(0.42) | 481 | 0.75(0.43) | 0.02(0.02) | 0.372 | 0.10 |
| Educational expectations | 486 | 0.61(0.49) | 479 | 0.58(0.49) | 0.03(0.03) | 0.223 | 0.12 |

## ANALYTIC MODEL SPECIFICATIONS

The primary impact study examines whether participation in the PGC-HS program impacts students staying in school (number of days attended during the ninth-grade year) and progressing through school (credits earned during the ninth-grade academic year). We assess program impacts using a regression equation that models outcomes as a function of treatment status, the baseline measure of the outcome variable, blocking variables, and other covariates. Although a straight difference-of-means/proportion approach would have provided unbiased estimates of the effect of the treatment intervention, a model-based approach is preferred with covariates because it increases the precision of those estimates.

The empirical model was estimated with an Ordinary Least Squares (OLS)/Linear Probability Model (LPM) regression model in Stata 15 (StataCorp, 2015). We model primary outcomes using the following empirical model:

$$Y_{Post} = \beta_0 + \beta_1 T + \sum(\beta_P X_P) + \varepsilon$$

Where:
$Y_{Post}$ is the outcome variable;

*T* is a dummy treatment indicator variable whose value equals 1 if the participant is randomized into the treatment group and 0 otherwise;

*X* is a *p* vector of baseline (i.e., measured prior to receiving intervention or exogenous to treatment) participant-level covariates as well as blocking variables to account for the variation in outcomes associated with these groups and to increase the precision of our impact estimates. These covariates include:

a) A pre-intervention measure of the outcome variable; variable is re-centered at the grand mean for analysis;
b) Age at baseline (continuous) as reported by study school; variable is re-centered at the grand mean for analysis;
c) Race/ethnicity of participant as reported by study school. Race is coded as a set of 4 - 1 = 3 dummy variables (each coded as 1 if they are of the specified race/ethnicity and coded as 0 otherwise); each of the variables is re-centered at the grand mean for analysis;
d) Gender of participant as reported by study school; a dummy variable (0 = male; 1 = female) that captures the differential effects associated with participants' gender; variable is re-centered at the grand mean for analysis;
e) English learner status – reported for each participant by study school; a dummy variable (0 = no ELL; 1 = ELL status) that captures the differential effects associated with participants' socio-economic disadvantage; variable is re-centered at the grand mean for analysis;
f) Special education (IEP) status as reported for each participant by study school; a dummy variable (0 = no IEP; 1 = has an IEP) that captures the differential effects associated with participants' academic disadvantage; variable is re-centered at the grand mean for analysis;
g) Block is an 8–1 vector of dummy variables to capture the effects of the eight schools (by cohort) that offered the intervention during the evaluation period. A student is coded as 1 if they attended a particular site during the specified year and 0 otherwise. Dummy variables are then re-centered at the grand mean for analysis.

$\beta_0$ is the intercept term, which represents the regression-adjusted mean of the outcome variable for the control group, with all other variables in the model held constant at 0.

$\beta_1$ is the parameter estimate of substantive interest and represents the adjusted mean difference in the outcome for those in the treatment condition.

We report the model-estimated difference between the treatment and control group ($\beta_1$), along with the model estimates for the treatment mean ($\beta_1 + \beta_0$) and control mean ($\beta_0$). Statistical significance is based on test statistics produced by Stata 15 for the coefficient $\beta_1$ using a two-tailed test, with $p < .05$.

## ASSESSMENT OF BASELINE EQUIVALENCE

We assessed baseline equivalence of treatment and control groups within each analytic sample by assessing the pre-intervention differences in important background characteristics and outcomes observed in data. To assess equivalence, we generated a model-based estimate of the difference

between treatment and control groups for the pre-intervention variables; the empirical model is a reduced form of the model used to estimate program impact (as specified in the Analytic Model Specifications section above). It is a reduced form because individual-level covariates are omitted. Separate models are run, and estimates provided, for each of the variables selected for baseline equivalence. Where the baseline variable is continuous, the model is estimated with OLS and the standardized difference is calculated using the Hedges' *g* formula; where the baseline variable is dichotomous, the model is estimated with a logistic regression model and the difference in the probability of the occurrence is calculated with the Cox Index formula.

### *CONTINUOUS VARIABLES*

The following model was used to produce estimates of baseline equivalence for continuous variables:

$$Y_{baseline} = \beta_0 + \beta_1 T + \sum(\beta_p Block_p) + \varepsilon$$

Where:

$Y_{baseline}$ is the baseline or pre-intervention measure that we use to establish baseline equivalence;

$T$ is a dummy treatment indicator variable whose value equals 1 if the participant is randomized into the treatment group and 0 otherwise;

$Block$ is an 8–1 vector of school blocking dummy variables that are coded as 1 if the participant attended the school during the specified school year and coded as 0 otherwise;

$\beta_0$ is the intercept term, which represents the adjusted mean value of the baseline measure for participants in the control sample, with all other variables in the model held constant at 0;

$\beta_1$ represents the adjusted (but not standardized) mean difference in the baseline variable between treatment and control participants.

Next, we computed the pooled standard deviation of the pre-intervention measures used to establish baseline equivalence. We used the following formula to compute the pooled standard deviation of the pre-intervention measure:

$$S_p = \sqrt{\frac{(n_t - 1)S_t^2 + (n_c - 1)S_c^2}{(n_t + n_c - 2)}}$$

Where $n_t$ and $n_c$ are the sample sizes, and $S_t$ and $S_c$ are the participant-level standard deviations for the pre-intervention measures for the analytic treatment and comparison groups, respectively. We produced separate calculations of the pooled standardized deviation for each variable used to establish baseline equivalence (as noted above).

We then produced the standardized difference of means using the formula for Hedges' *g*:

$$g = \frac{\beta_1}{S_p}$$

Where $\beta_1$ is the adjusted mean difference in the variable selected to establish baseline equivalence for the treatment and comparison groups (calculated in Step 1), and $S_p$ is the pooled standard deviation (calculated in Step 2).

### *DICHOTOMOUS VARIABLES*

According to the What Works Clearinghouse, "The effect size measure of choice for dichotomous outcomes is the Cox Index, which yields effect size values similar to the values of Hedges' $g$ that one would obtain if group means, standard deviations, and sample sizes were available."[17] Following this guidance, we used the Cox index to estimate baseline equivalence for dichotomous baseline covariates using the following formula:

$$d_{Cox} = \left[ ln\left(\frac{p_t}{1-p_t}\right) - ln\left(\frac{p_c}{1-p_c}\right) \right] \Big/ 1.65$$

Where $p_t$ and $p_c$ represent the probability of occurrence of the event (or characteristic) within the treatment and comparison groups, respectively.



[17] What Works Clearinghouse Procedures Handbook, Version 4.1.