Impact Evaluation Report

Baboquivari Unified School District
Wisdom Project

November 2016

The Policy & Research Group
8434 Oak Street
New Orleans, LA 70118
www.policyandresearch.com
504.865.1545

The Policy & Research Group

Table of Contents

Executive Summary

The Baboquivari Unified School District (BUSD) implemented its Wisdom Project from 2012-2015 through the Investing in Innovation (i3) grant from the United States Department of Education. The Wisdom Project focuses on improving academic achievement and increasing students' readiness for college. BUSD contracted with The Policy & Research Group to conduct the evaluation of the Wisdom Project. The purpose of this report is to describe the evaluation of the Wisdom Project and provide results from the impact study at the project's conclusion.

The Wisdom Project offered a combined set of college readiness programs to students at Baboquivari Middle School (BMS) and Baboquivari High School (BHS). At both schools, the Wisdom Project implemented a comprehensive schoolwide college readiness strategy that includes teacher training, improvements in technology, and various seminars and trainings to help students prepare for higher education. This strategy was available to all students at both schools, though some activities for middle school students differed from those for high school students. Along with the schoolwide strategy, the Wisdom Project also offered an Advancement via Individual Determination (AVID) elective course for some students at both BMS and BHS. The AVID elective course teaches organizational and study skills, offers college tutoring, and provides motivational activities intended to make college seem more attainable.

At both BMS and BHS, some students are only exposed to the schoolwide (college readiness) program. Students who are enrolled in the AVID course are also exposed to the schoolwide college readiness program. As a consequence of this overlap, we are constrained in how we may assess the impacts of these programs on student outcomes. Since there are some students who are only exposed to the schoolwide program (and not AVID), we are able to assess the impact of the schoolwide program on students who are not selected to receive AVID. On the other hand, since AVID students are necessarily exposed to both AVID and the schoolwide program, we cannot assess the impact of the AVID program by itself. Instead, we must assess the impact of the combined AVID and schoolwide strategies together. We refer to this compound intervention as the "AVID-Core-Plus" intervention.

In this *Impact Evaluation Report* we present findings of two separate studies. We evaluate the impact of the compound AVID-Core-Plus intervention on student outcomes in one study, and, in the other, we evaluate the impact of the schoolwide program on student outcomes. Because students who are enrolled in the AVID course are also exposed to the schoolwide program, our evaluation of program impacts on these students is unable to isolate any effects of the AVID course alone; for these students, we evaluate the combined effects of participation in both programs. For both studies, our research questions ask whether or not the intervention impacts changes in students' math and reading standardized test scores and whether or not it impacts participating students' college readiness.

The primary study assesses the impact of the AVID-Core-Plus intervention – the combination of the AVID elective course and the schoolwide intervention – on participating students' academic achievement in math and reading and on their college readiness. Academic achievement is operationalized as the relative change in students' standardized math and reading test scores over the course of the intervention period, using rank-based z-scores. College readiness is operationally defined as a scale score calculated from student responses to questions about their academic preparedness, college knowledge, and college admissions preparedness. We estimate these impacts by comparing these outcomes for students in the treatment group with those from a group of comparison students who are similar in observable characteristics. To assess program impact on academic achievement, we use multi-level difference-in-difference models to determine if students who received the intervention improved their test scores at greater rates than did similar comparison students who did not receive the intervention. To assess program impact on students' college readiness, we use a multi-level regression model to determine if students in the 11th and 12th grades who received the intervention were more college-ready than were similar comparison students in the same grades who did not receive the intervention.

The secondary study assesses the impact of the schoolwide intervention on the same academic achievement and college readiness outcomes. We estimate these impacts by comparing these outcomes for students in the treatment group with those from a group of comparison students who are similar in observable characteristics. The analytic methods used in this study are identical to those employed in the first study.

Key Findings

Primary Study: AVID-Core-Plus Intervention

- The AVID-Core-Plus intervention had no noticeable effect on academic achievement in math or reading. In math, students in both the treatment and comparison groups experienced a decrease in relative math test performance during the study period. In reading, relative performance of students in the treatment group increased while relative performance of comparison group students decreased. In each of these cases, the difference is not statistically significant.

- The AVID-Core-Plus intervention had a positive impact on college readiness. We find that students exposed to the AVID-Core-Plus intervention are better prepared for college than students in the comparison group, as measured by the College Readiness Index. At the end of the 11th and 12th grades, the mean College Readiness Index scores for students in the treatment group are significantly higher than those for students in the comparison group.

Secondary Study: Schoolwide Intervention

- The schoolwide intervention did not have a positive impact on academic achievement in math or reading. Results from our analyses show that, as compared to the comparison group, students in the schoolwide intervention demonstrate a greater decline in math and reading test performance over the course of the study. The difference in the relative decline is statistically significant.

- The schoolwide intervention had no discernable impact on college readiness. Results of this study indicate that the schoolwide intervention did not cause any change in students' college readiness, as measured by the College Readiness Index. Students in the treatment and comparison groups report College Readiness Index scores that are nearly identical.

Introduction

The Baboquivari Unified School District (BUSD) implemented its Wisdom Project from 2012-2015 through the Investing in Innovation (i3) grant from the United States Department of Education. The goal of the i3 grant is to test innovative educational practices to assess their impact on improving student achievement or student growth, closing achievement gaps, decreasing dropout rates, increasing high school graduation rates, and increasing college enrollment and completion rates. The Wisdom Project focuses on improving academic achievement and increasing students' readiness for college.

The i3 grant requires a rigorous evaluation of each grantee's project. BUSD contracted with The Policy & Research Group to conduct the evaluation of the Wisdom Project. The purpose of this report is to describe the evaluation of the Wisdom Project and provide results from two impact studies of Wisdom Project strategies.

Structure of the Report

In this *Impact Evaluation Report,* we first present an overview of the Wisdom Project, which was implemented in BUSD from 2012-2015. We then list the research questions that guide the evaluation of the impact of the program on academic outcomes and college readiness of participating students. We outline the study design, outcome operationalization, eligibility criteria, matching process, and analytic methods, and we provide a description of the baseline equivalence of our treatment and comparison samples for each outcome. We then provide our results and a discussion of our interpretations of these results, as well as the limitations of our analytic methods. This report also contains a number of appendices that provide additional details on information presented in the body of the report. Appendix A provides details on our analytic methods. Appendix B describes transformations of the standardized test scores used as our academic outcomes. Appendices C and D provide the item rationale and scoring protocol for the College Readiness Index. Appendix E provides details on data management for datasets received from the Arizona Department of Education (ADE). Appendix F provides descriptive and diagnostic statistics of the analytic samples and the baseline equivalence of the samples. Finally, Appendix G provides model specifications for our analytic results.

Program Description

<u>The Wisdom Project</u>

The Wisdom Project is a high school completion and college- and career-readiness initiative that serves students in grades 4-12 in BUSD. The Wisdom Project aims to improve academic achievement, college readiness, high school completion, and college enrollment among participating students in BUSD.[1] In 2012, the Wisdom Project received an i3 grant to implement a set of programs as an innovative educational practice. As a part of this effort, the Wisdom Project introduced the Advancement via Individual Determination (AVID) elective course for high school and middle school students, along with a schoolwide college-readiness program at both Baboquivari Middle School (BMS) and Baboquivari High School (BHS).

BHS and BMS first offered the AVID elective course in 2012 to students in grades 7 through 10, and it was expanded over the next two years to include students in grades 6 through 12. The AVID elective course was implemented as a part of the schoolwide AVID Secondary program, in which all teachers are trained to use AVID strategies in their classes to develop students' academic skills. The AVID elective

[1] The Wisdom Project also aims to improve technology use in instruction and postsecondary planning to improve participants' comfort with and use of technology. Originally, the Wisdom Project also aimed to serve young adults at risk for dropping out of school or those who had already left school who re-enrolled in an alternative education program in BUSD. This evaluation study does not assess program impacts on these outcomes.

course specifically aims to improve middle and high school students' college enrollment and preparedness by teaching critical thinking, teaching organizational and study skills, providing academic assistance from peers and college tutors, and providing motivational activities intended to improve students' perceptions of their ability to attend college. AVID describes the target group of students for the AVID Secondary program as those who are "the first in their families to attend college, and [who] come from groups traditionally underrepresented in higher education."[2]

In addition to the AVID elective course, the Wisdom Project also implemented a schoolwide college readiness program at BHS and BMS. This program added AVID study techniques to non-AVID courses, updated classroom technology to include Infinite Campus and Promethean/SMART boards, and offered a series of programs and trainings for students and their parents on college and career readiness. These programs include community college dual enrollment classes for high school students, career exploration and college preparation programs for middle and high school students during the summer and the school year, and workshops on postsecondary education for students and their parents. These college readiness activities were offered to all students at BHS and BMS; they were not limited to students enrolled in the AVID elective course.

At the outset of the intervention, Wisdom Project staff hypothesized that these activities would lead to changes to teaching methods, student course choices, student learning methods, and student self-efficacy, knowledge, beliefs, attitudes, engagement, and intentions related to college preparedness. They further hypothesized that the short-term outcomes of these programs would include improved student achievement and improved college readiness. The long-term outcomes (not measured by this evaluation) are expected to include increased postsecondary applications, increased postsecondary acceptance and enrollment, increased postsecondary attendance rates, and, ultimately, increased postsecondary completion, certification, and graduation. Although the program aims to impact college attendance, this not a feasible outcome to measure over a three-year period, given that the grant ended when the oldest cohort of students in the intervention were at the end of 12th grade.

Evaluating the Wisdom Project Strategies

The Wisdom Project began to offer programming in 2012. Schoolwide college readiness activities were initiated at BMS and BHS in 2012 for students in grades 6 through 12, but the AVID elective course was offered to a narrower group of students. Beginning in 2012, the AVID elective course was offered to selected students at BHS and BMS in grades 7 through 10. In 2013, it was offered to students in grades 6 through 11, and, in 2014, it was offered to students in grades 6 through 12.

Both Wisdom Project programs were introduced at both schools at the same time. All students in both schools were exposed to the schoolwide college readiness program, and students who were selected into the AVID course were offered the AVID course and the schoolwide program simultaneously. As a result of this, the evaluation team is able to isolate the impact of the schoolwide program (non-experimentally), but we are unable to construct a design to isolate the effect of the AVID program alone.[3] Since every student who was exposed to the AVID program was also exposed at the same time to the schoolwide college readiness program, there is no valid way to identify the effects of AVID without also including those of the schoolwide college readiness program. As such, for those exposed to AVID and the

[2] Retrieved from http://www.avid.org/what-is-avid-secondary.ashx on July 21, 2016.

[3] In the impact evaluation, we estimate program impacts by comparing students exposed to the schoolwide and AVID-Core-Plus programs at BUSD to students at different (but similar) schools. We compare BUSD students to students in other school districts instead of comparing AVID students in BUSD with non-AVID students in BUSD because AVID guidelines dictate that students who are selected into AVID are necessarily different from those who are not. Contrasting academic outcomes for these two groups as a measure of program impact would likely insinuate bias. Since the background characteristics that differentiate the two groups also can be expected to explain some of the variation in outcomes, the resulting impact estimates would not just represent the impact of the program; rather, it would include some (unknowable) combination of background influence and the impact of the program. Since the college readiness program was offered to all students in the school, an internal (i.e., within BUSD) comparison is not practicable.

schoolwide strategies, we estimate the combined effects of both as a single intervention. Below we describe the two evaluation studies.

Primary Study: AVID-Core-Plus Intervention
Our primary study assesses the impact of the AVID elective course plus the schoolwide college readiness program on two student outcomes: relative improvement in math and reading achievement, and college readiness. We refer to the combination of the AVID elective course and the schoolwide college readiness activities as the "AVID-Core-Plus" intervention. For this study, we identify the treatment group as students who were enrolled in the AVID elective course for at least one year.[4] The comparison group is a selected group of similar students who attend schools in other districts in Arizona that are comparable to BUSD in observable background characteristics at baseline. We assess program effects as the relative difference in treatment and comparison group performance in standardized test scores from pre- to post-intervention. We also assess whether or not 11th and 12th grade students in the treatment group demonstrate greater college readiness than do members of the comparison group.

Secondary Study: Schoolwide Intervention
Our secondary study assesses the impact of participation in the schoolwide college readiness program on the same student outcomes: relative improvement in math and reading achievement, and college readiness. In this report, we refer to this program as the "schoolwide" intervention. For this study, we identify the treatment group as students who were enrolled in BHS or BMS for at least one year while this program was active. The comparison group is a selected group of similar students who attend schools in other districts in Arizona that are comparable to BUSD in observable background characteristics at baseline. We assess program impact as the relative difference in treatment and comparison group performance in standardized test scores from pre- to post-intervention. We also assess whether or not 11th and 12th grade students in the treatment group demonstrate greater college readiness than do members of the comparison group.

Impact Evaluation Design

This evaluation investigates the impact of the Wisdom Project components on participating students' academic achievement and college readiness. Outcome and student characteristic data are collected from state standardized test scores and from questionnaires collected directly from students. The study is a non-experimental or Quasi-Experimental Design (QED) that relies on statistical techniques to improve the equivalence of the treatment and comparison groups so that we may make valid inferences of program impact.[5] A non-experimental design relies on procedures other than randomization to create a treatment and comparison contrast. In this case, to create this contrast, we compare students in BUSD who receive the interventions to similar students (in observable characteristics) in Arizona but outside BUSD who do not receive the interventions. Although this method has limitations (discussed below), the evaluation team believes it is the most rigorous available given the constraints imposed in the planning stages.[6]

A QED attempts to draw causal inferences about program impact by comparing students who receive the intervention (treatment group) to other students who are similar in observed background characteristics but who do not receive the intervention (comparison group).[7] The study contrasts students who receive

[4] Because the college readiness intervention was implemented school-wide, these students were also exposed to the college readiness program.
[5] See Appendix A for a technical account of the methods used for these analyses.
[6] We apply this technique in part because random assignment of students into treatment and control conditions was deemed impracticable by program leadership in the planning stage of the evaluation.
[7] For both outcomes, students are matched by grade. For the academic outcomes, we match students based on their pre-intervention test score, gender, Native American status, English language learner status, special education status, and free and reduced-price lunch status. For college readiness, we match students based on parents' education, household resources, Native American status, gender, age, and living-with-biological-parents indicator.

the interventions with students from other districts outside BUSD who have received otherwise typical educational programming in an attempt to isolate the effect that the interventions have on outcomes of interest. By comparing groups of students who are equivalent in observed characteristics, the evaluators attempt to reduce the set of alternative factors that could influence outcomes and reduce the accuracy of our impact estimates. To maximize the equivalence of the two groups, the study employs a multi-stage matching procedure that first pre-screens for similarity at the district level and then matches at the individual student level.[8]

In addition to the matching procedures, the study also employs statistical analysis methods that attempt to mitigate any remaining and unobserved differences in the treatment and comparison groups. These techniques are explained briefly below and in more technical detail in Appendix A. Gains in academic improvement (in math and reading) are estimated by way of a difference-in-differences (DID) method that is common in program evaluation and education research.[9] College readiness outcomes are assessed at posttest only. We use multi-level empirical models to produce impact estimates, controlling for relevant covariates at the individual and school level.

We use the DID method to estimate the impact of the program on students' academic achievement. Using the DID method, we assess how students' academic achievement changes over time, and whether or not this difference over time differs based on membership in the treatment or comparison group. The DID method models how participation in the intervention impacts academic achievement over time, along with the effects of any relevant covariates (e.g., gender, Native American status).

We use linear regression to estimate the impact of the two interventions on college readiness, measured at posttest by the College Readiness Index. We assess the relationship between the outcome variable, college readiness, and treatment/comparison group status and relevant covariates (e.g., gender, Native American status), and whether or not college readiness differs between treatment and comparison students after controlling for the covariates. We assess this outcome at posttest only; although, we cannot measure changes in college readiness over time, we match treatment and comparison students in order to minimize differences between these groups that could introduce bias.

Research Questions

The evaluation includes six research questions that investigate academic and college readiness outcomes for both the AVID-Core-Plus intervention and for the schoolwide intervention. We list these questions first by study priority and then by outcomes investigated.

---

[8] The chief limitation of the QED is that the researcher cannot ever be completely sure that the treatment and comparison groups are in fact similar in terms of characteristics that go unobserved or are practically unmeasurable. Although it would be practically impossible to determine if the groups are not similar and if that difference has an effect on the outcomes that we measure, the resulting estimate of impact will conflate this factor (that we aim to exclude) with the program impact (that we aim to identify). This is known as bias, and it will either result in a higher or lower estimate of program effectiveness than actually exists. This limitation exists for all studies that are unable to randomly assign study participants. An additional limitation of this study is that the treatment condition is only implemented in a single district. This is sometimes called an $n = 1$ confound, and it means that the evaluators are unable to disentangle the effects of the intervention itself from the effects of the district. Despite these imposed limitations, however, the evaluators believe that design employed in this evaluation represents the most rigorous possible method to estimate the impact of the program on participating students.

[9] E.g., Angrist and Pischke, 2009; Schlotter et al., 2010; Antonakis et al., 2011; Murnane and Willett, 2011.

Primary Research Questions

Academic Outcomes

1. Do students who receive the AVID-Core-Plus intervention demonstrate more relative improvement in *mathematics achievement* than equivalent students who attend similar schools that have not received the intervention?

2. Do students who receive the AVID-Core-Plus intervention demonstrate more relative improvement in *reading achievement* than equivalent students who attend similar schools that have not received the intervention?

College Readiness

3. Do students who receive the AVID-Core-Plus intervention demonstrate higher levels of self-reported college readiness in the final year of the intervention than equivalent students in the comparison district that have not received the intervention?

Secondary Research Questions

Academic Outcomes

1. Do students who receive the schoolwide intervention demonstrate more relative improvement in *mathematics achievement* than equivalent students who attend similar schools that have not received the intervention?

2. Do students who receive the schoolwide intervention demonstrate more relative improvement in *reading achievement* than equivalent students who attend similar schools that have not received the intervention?

College Readiness

3. Do students who receive the schoolwide intervention demonstrate higher levels of self-reported college readiness in the final year of the intervention than equivalent students in the comparison district that have not received the intervention?

Methods

Identifying Treatment Students

Here, we outline study definitions for identification into the treatment group, and, in the next section, we outline the requirements and procedures for how students are selected into the comparison group. We refer to the group of students selected into the treatment group as the treatment sample and the group of students matched for comparative purposes as the comparison sample. The combined group of treatment and comparison groups we refer to as the analytic sample.

Primary Study

Academic Achievement Outcomes: To be eligible for inclusion into the treatment samples for the analysis of academic outcomes, a student must have attended a study school (BHS or BMS) during the study period and must have been identified by BUSD as a participant in the AVID-Core-Plus intervention for at least one year during the study period.[10] The study period includes students in grades 7 through 10 in the 2012/13 school year, grades 6 through 11 in the 2013/14 school year, and grades 6 through 12 in the 2014/15 school year.[11] Students must have complete pre-intervention and post-intervention test data.[12] The pre and posttests must be linkable (by student ID) and in the same subject.

To be eligible for selection into the comparison group – that is, the pool of students from which we matched comparison students to students in the treatment group – students must attend a set of pre-screened comparison school districts at the same time as treatment students and have all relevant data. Administrative data from ADE must indicate that a student attended a school in a comparison school district during the study period and was enrolled in the same grades as treatment students during the study period, and the student must have complete pre-intervention and post-intervention test data.[13] The pre and posttests must be linkable by student ID and in the same subject. Students must not be identified as having participated in the AVID-Core-Plus intervention.

College Readiness: To be eligible for inclusion into the treatment group for the analysis of program impacts on college readiness, 11th and 12th grade students at BHS must have completed the CRQ and indicated on the questionnaire that they participated in the AVID-Core-Plus program. To be eligible for the comparison group, 11th and 12th grade students at the pre-screened comparison high school must have completed the CRQ.

<u>Secondary Study</u>

Academic Achievement Outcomes: To be eligible for inclusion into the schoolwide treatment group for the analysis of academic outcomes (math and reading), administrative data provided by BUSD must indicate that a student attended a school where the intervention was offered (BHS or BMS) during the study time period. Students must have complete pre-intervention and post-intervention test data, which must be linkable (by student ID) and in the same subject.[14] The pretest is a standardized test taken before the student is eligible to participate in the schoolwide intervention. Pretest data include standardized tests taken by students in grades 4 through 11 during the 2011/12 school year, or by students in grades 4 and 5 in the 2012/13 school year, or in grade 5 in the 2013/14 school year. Posttest data include standardized tests taken by students in grades 6 through 12 in the 2012/13, 2013/14, or 2014/15 school years. The most recent available standardized test taken by a student identified as having attended BHS or BMS during the study period is used as that student's posttest.

To be eligible for inclusion into the pool of students from which we select the comparison group for the secondary study, students must attend a set of pre-screened comparison school districts at the same time as treatment students and have all relevant data. Administrative data from ADE must indicate that a student attended a school in a comparison school district during the study time period and was enrolled in the same grades as treatment students during the study period, and they must not indicate that these students attended an intervention school (BHS or BMS) for any portion of the time when the intervention was offered. The intervention was offered to students in grades 6 through 12 during the 2011/12 through 2014/15 school years. Because comparison group students are matched to treatment group students in

[10] We determine if a student attended a study school on the basis of administrative data provided to PRG by ADE. We received a dataset from BUSD in 2013 identifying students who participated in the AVID-Core-Plus program for each year that it was offered. Students who participated in the program for less than one year are not included in the treatment group.
[11] See Tables A2 and A3 in Appendix A for an illustration of the grades and years when the intervention was offered.
[12] 13 students in the AVID-Core-Plus treatment group were missing math test data, and 11 students were missing reading test data.
[13] For each year of the evaluation, we received data from the Arizona Department of Education (ADE) with students' standardized math and reading test scores. Some students in the dataset were missing either a math or a reading test score: 33 students in the comparison group had a math score but not a reading score, and 124 students had a reading score but not a math score.
[14] In the schoolwide treatment group, 12 students had a math score but not a reading score, and 13 had a reading score but not a math score.

each grade in each year of the intervention, in order to be a part of the pool from which the comparison group is selected, students must have complete pre-intervention and post-intervention test data, which must be linkable (by student ID) and in the same subject, in the grades and years noted above.

College Readiness: To be eligible for inclusion in the schoolwide treatment group, 11th and 12th grade students must have completed the CRQ at BHS in spring 2015 and indicated on the questionnaire that they did not participate in the AVID-Core-Plus program. To be eligible for the comparison group for college readiness, 11th and 12th grade students at the pre-screened comparison high school must have completed the CRQ.

Selecting Comparison Students that Match the Treatment Sample

Separate analytic samples were constructed for each of the six outcome research questions. The initial process involved first identifying, or pre-screening, potential pools of students for matching by identifying school districts that were similar to the treatment district in terms of five key characteristics. From these samples, we select comparison students who are closely matched to treatment group students based on relevant characteristics using the procedures described below. Specifically, students in the comparison condition are matched (based on relevant background characteristics) to students in the treatment group for each outcome in each year of the study. With these matching procedures, we seek to maximize the equivalence of the treatment and comparison groups in terms of key observed variables by selecting only those comparison students who most closely approximate the treatment students in terms of those variables. We describe the matching procedures below and in detail in Appendix A.

Stage 1: Screening of Similar Districts

To create a credible match for treatment students, we first identified a set of school districts that are equivalent to our treatment district in key background characteristics.[15] Baseline (i.e., pre-intervention) characteristics that were considered are: (1) proportion of the school district that is Native American; (2) urbanicity of the district; (3) past academic performance of the school district; (4) population of the district; and (5) economic indicators. A preliminary analysis of the school districts in Arizona indicated that districts are either predominantly Native American or not. When districts are sorted by this proportional distribution, an evident cut-point occurs in districts that are around 90% Native American.

Based on these criteria, we selected five school districts that are, in composition, comparable to BUSD. Each of the districts is predominantly Native American, rural, remote/distant, and rated similarly by ADE in terms of their academic performance in 2011. Additionally, the districts appear to be alike in terms of median household income and size (population). For college readiness outcomes, one comparison district agreed to complete the CRQ; the district is comparable to BUSD on relevant district-level characteristics. See Table A1 in Appendix A for data on selected and non-selected comparison districts.

Stage 2: Individual-Level Matching

In order to select comparison students who are closely matched to treatment group students, we first select the relevant characteristics by which to match comparison group students to students in the treatment group. Then, using these variables, we calculate the "distance," or similarity, between students in the treatment group and potential comparison students. Finally, for each year of the study, we match treatment group students in each grade to similar students in the comparison group in the same grade.

[15] This is consistent with the guidance offered by Song and Herman (2010) and Cook, Shadish, and Wong (2008). Song, M. & Herman, R. (2010). Critical Issues and Common Pitfalls in Designing and Conducting Impact Studies in Education: Lessons Learned from the What Works Clearinghouse (Phase I). *Educational Evaluation and Policy Analysis*, 32, 351-371.
Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiment and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27(4), 724–750.

Variables to Include in Match
For each outcome (academic achievement in math, academic achievement in reading, and college readiness), we identify relevant characteristics to determine the similarity between treatment and comparison students. For matching our academic outcomes samples, we include a pre-intervention measure of the outcomes (math and/or reading AIMS pre-test scores) as well as all other variables that we expect are highly correlated with the outcome. For matching our college readiness outcomes samples, we include a measure that is, theoretically, highly correlated with the outcome (parents' education) as well as all other available variables that we expect are highly correlated with the outcome.

Academic Achievement Outcome Matching Variables

- Gender
- Native American indicator
- English language learner status indicator
- Special education status indicator
- Free and reduced-price lunch status indicator
- Pre-intervention test (pre-test) score(s)

College Readiness Outcome Matching Variables

- Gender
- Native American indicator
- Age
- Living with both biological parents
- Household resources (Measure of SES)[16]
- Parents' education (Measure of SES)

Distance Metric Selection
The evaluation team decided to use Euclidean distance matching with propensity score calipers as our distance metric. Euclidean distance matching calculates the "distance" between a student in the treatment group and potential comparison students, based on how similar they are (i.e., the "distance" between the two participants) on the full set of matching variables. We select Euclidean distance matching because it is recommended for small samples and a limited number of covariates. Further, since there is a single covariate that is more predictive of the outcome variable than the others (the pretest for academic outcomes; parental education for college readiness), Euclidean distance matching is ideal because it can incorporate this primacy in the distance calculation by weighting that covariate more than the others.

Matching Procedures
Finally, matching is conducted for each outcome for students in each grade for each outcome. There were six cohorts of students that moved through the AVID-Core-Plus and schoolwide interventions, based on their grade level and the year they were first exposed to the intervention. Tables A1 and A2 in Appendix A describe cohort progression through the AVID-Core-Plus and schoolwide interventions.

In each year of the intervention, students who are eligible for inclusion into the treatment group are matched with similar students in the same grade from the pool of eligible comparison students. Within each cohort set, we calculate a distance between all possible treatment-comparison pairs and then select comparison cases on the basis of a 1:1 closest match, without replacement, for each treatment/comparison pair that minimizes the Euclidean distance. That is, matching is done by identifying the closest comparison match within each cohort for each treatment case, removing that pair, and then matching the

[16] See College Readiness Survey Q.9 in Appendix E.

next closest treatment-comparison pair. This routine continues until all eligible treatment cases have been matched with a comparison case within each cohort.

Because we have decided to use propensity score calipers, only those matched cases whose absolute difference in propensity scores do not exceed .25 of a standard deviation are eligible to be included in the analytic sample. We estimate propensity scores using logistic regression and our selected set of covariates for each outcome (academic achievement and college readiness). For academic outcomes, we estimate propensity scores using the following variables: gender, English language learner status, special education status, free and reduced lunch status, and pre-intervention test score for same subject as the outcome.

The matching procedure differed for the college readiness outcome samples because of differences in the availability of baseline data. We were not able to gather student-level data prior to the intervention, nor are we able to reliably link data obtained in the questionnaire with other student-level data (i.e., test data). As such, we match individuals based on data gathered in the *College Readiness Questionnaire* instrument that was administered in the final year of the intervention (2014-15). Any student who completed the questionnaire and indicated that he/she has participated in the AVID-Core-Plus program is considered a member of the core-plus treatment group. All other respondents who completed the questionnaire at BHS are considered schoolwide participants. Students who completed the questionnaire at the comparison high school are considered eligible for matching for either (or both) the primary and secondary studies. We estimate propensity scores for the college readiness samples using the following variables: gender, age, grade, a lives-with-both-parents indicator, a household resources indicator, and a parent education indicator.

Outcome Measures

<u>Academic Outcomes</u>

We measure academic outcomes using rank-based z-score transformations of AIMS and AZMERIT standardized test scores in reading and mathematics. The AIMS test was replaced by the AZMERIT test in 2015, but one or the other was required for students in grades 3 through 8 and grade 10 during the study period. Student-level test data (scaled scores) were received from ADE for all students for each year of the study and for one year prior to the start of the intervention for the initial cohorts. For each grade in each year of the study, we transform the scaled scores to rank-based z-scores according to the procedures outlined in the *Model Specifications* section of Appendix A. We elect to transform scores from their scaled score for two reasons: 1) the high school AIMS test (which was taken by students in years prior to 2015) is not on the same scale as the test for grades three through eight; and 2) ADE transitioned from the AIMS test to a new standardized test, the AZMERIT, during the final year of the intervention study. Re-scaling student-level data to a common metric is an appropriate analytic method to deal with these inconsistencies.[17]

Ranked z-scores are a transformation of scaled test scores that indicate the performance of a student relative to the performance of all other students in the state as a whole on a given standardized test. We calculate these transformations for all students for each test in each academic year. A rank-based z-score of 0 corresponds with an average score statewide on that test; a score of 1 indicates a score that is one standard deviation above the average; and a score of -1 indicates a score that is one standard deviation below the average. For example, a 6^{th} grade student in 2011 who has a math z-score of 0 has a score that is equivalent to the mean for the entire state for that test in that year (a raw score of 413). If that student

[17] This is consistent with guidance offered by May et al., (2009). May, H., Perez-Johnson, I., Haimson, J., Sattar, S., and 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.

has a z-score of 1, then that student's score is a full standard deviation higher than the mean test score for that test in that year (a raw score of 465).[18]

College Readiness Outcomes

College readiness outcomes are measured using the College Readiness Index (CRI), which is a summative index score of participant responses to a series of items included in the *College Readiness Questionnaire* (CRQ). The CRI is a 17-item index that quantifies the respondent's self-reported college readiness. The index includes items that measure three latent variables – academic preparedness, college knowledge, and college admissions preparedness – that have been identified as predictive of being prepared for college (see item rationale in Appendix C). Each item/concept included in the index is scored according to the scale scoring protocol provided in Appendix D. The analytic variable represents the total index score for each student. The CRI operationalizes the domain of college readiness, which is the degree to which an individual is prepared for and able to attend college.

Items in the index measure academic preparedness (e.g., content knowledge, academic skills, and academic performance), college knowledge (e.g., understanding of the college admissions process), and college admissions preparedness (e.g., students have taken steps necessary to enroll in college). Face validity is established on the basis that the requirements listed within the index are identified in the literature as a predictor of college attendance and/or as a basic requirement for college eligibility, acceptance, and/or enrollment (see Appendix C for justification of item inclusion and measure construction). We use Cronbach's alpha to determine the internal consistency (reliability) of all the items included in the scale based on a pre-administration of the instrument in April 2013.

Data Collection

Academic Outcomes

PRG entered into a data sharing agreement with ADE to receive statewide math and reading scores and demographic data for all students in grades 4, 5, 6, 8, and 10 from 2011-2015. We requested and received data from ADE in spring 2015 for testing data from 2011-2014, and again in fall 2015 for testing data for the 2014-15 school year. The datasets ADE provided are organized by testing year, with unique observations for math scores and reading scores. Each observation includes the following demographic data: race, gender, special education status, English language learner status, and free/reduced lunch status. We merged the individual-level data from year to year using masked student ID numbers provided by ADE.[19]

College Readiness Outcomes

The CRQ was administered to all students at BHS and a single comparison high school during the spring of 2015; however, only 11th and 12th grade students are included in the college readiness outcome. PRG analysts administered the CRQ to all 11th and 12th grade students who were in attendance at BHS on May 5, 2015. BHS staff proctored questionnaires within one week for students who were absent during the initial administration. Students completed paper questionnaires, which were returned to PRG's office for data entry and analysis. At BHS, 82 11th and 12th grade students completed a questionnaire.

PRG identified a single high school in a comparison school district that agreed to complete the CRQ. Part of the agreement stated that the district would remain anonymous. PRG staff administered the CRQ to all 11th and 12th grade students in attendance at the comparison district's high school on May 6-7, 2015; school staff proctored questionnaires within one week for students who were absent during the initial two-

[18] Tables B1 and B2 in Appendix B describe how z-scores correspond to standardized test scores for each year of the project.

[19] We requested and received datasets from ADE in each year of the study (2011-2015), which included statewide math and reading scores and demographic data for all students in grades 4-6, 8, and 10. ADE de-identified all data before sending to PRG. For details on data management, see Appendix E.

day administration. In the comparison district, 78 11th and 12th grade students completed a questionnaire. Procedures for administration were identical at the treatment and comparison schools.

Analysis

The *Model Specifications* section in Appendix A outlines the analytic approach for the six research questions. Below, we provide an overview of the analytic approaches for providing an empirical response to the six research questions across the two outcome domains, academic achievement and college readiness.

Academic Outcomes

To answer the first two research questions in each study, we construct an empirical model that estimates the effects of each intervention (AVID-Core-Plus and schoolwide) on reading and mathematics outcomes. We estimate these impacts for a pooled sample of students who have been exposed to the intervention at some point in grades 6, 7, 8, and 10. Students may have been exposed to the intervention for as many as three years or as few as one year. Our objective is to assess the extent to which exposure to the intervention impacts relative changes in students' rank-based z-scores within each study period. We do this by constructing and analyzing the results from four separate models: for our primary study, we assess the impact of the AVID-Core-Plus program on achievement in reading and math, and, for our secondary study, we assess the impact of the schoolwide program on achievement in reading and math.

The models we use are multi-level difference-in-difference (DID) models that provide an estimate of how much students' rank-based z-scores change within each study period (i.e., from the pretest to the posttest), and the extent to which this change differs between the treatment and comparison groups. We regress rank-based z-scores on predictors of interest – treatment group, time, and their interaction – and a vector of relevant covariates (demographic variables). The interaction between time and the treatment group is our variable of interest – this variable indicates whether or not the treatment group experienced a significant change in test scores relative to the comparison group.

College Readiness Outcomes

To answer the third research question in each study, we construct an empirical model that estimates the effects of the AVID-Core-Plus intervention and the schoolwide intervention on college readiness of students in the 11th and 12th grade in the analytic sample. We estimate these impacts on two cohorts of students (those who were in the 11th or 12th grade in the 2014/15 school year) who have been exposed to the intervention at some point in grades 9 through 12. Students may have been exposed to the intervention for as many as three years or as few as one year. Our objective is to assess the extent to which exposure to the intervention impacts students' college readiness, indicated by the CRI score. We construct and analyze the results from two separate models: our primary study assesses the impact of the AVID-Core-Plus intervention on college readiness, and our secondary study assesses the impact of the schoolwide intervention on college readiness.

We use linear regression to estimate the impact of the program on students' college readiness. We use a multi-level model to regress the outcome variable (college readiness) on the predictor of interest (treatment/comparison group status), along with a vector of relevant covariates (demographic variables and parental education pre-intervention).

Baseline Equivalence

For each research question in each study, our aim is to produce an unbiased estimate of the effect of the intervention on the relevant student outcome. Because we cannot randomly assign participants to the treatment and comparison groups, we rely on matching procedures to maximize the equivalence of students in the treatment and comparison samples based on relevant background characteristics. The validity of our impact estimates are entirely dependent on the effectiveness of the matching, which is to say the actual equivalence of the treatment and comparison samples on these background characteristics. Although we cannot assess balance on unobserved characteristics (the principal weakness of any quasi-experimental design), we can assess the similarity between the treatment and comparison groups on a range of variables that are observed before participation in the intervention.[20] To test for baseline equivalence, we produce standardized measures of difference between the two samples for each analytic sample used for each research question.[21] Please see the *Methods for Baseline Balance Testing* section in Appendix A for detailed information and the specific models used to establish baseline balance.[22]

Academic Outcome Samples

The samples of students for academic outcomes are well-balanced across observed characteristics. Treatment and comparison groups do not differ by more than .25 standardized units for any observed characteristic for both primary and secondary study samples. Please see Tables F1-F4 in Appendix F for detailed results on balance statistics.

College Readiness Outcome Samples

The analytic sample for the primary study on college readiness outcomes is well-balanced across observed background characteristics. Although the treatment and comparison samples do differ by .25 standardized units in Native American status, this difference is due to the low level of variation in this characteristic, since nearly all students are Native American; these groups differ by only one student who does not identify as Native American. For all other characteristics, differences between the treatment and comparison groups are less than .25 standardized units. Nearly all students identified as Native American – 97% of the treatment group identified as Native American versus 100% of the comparison group.

The analytic sample that is used to respond to the single secondary study question on college readiness impacts is slightly less balanced. The groups differ by .33 units for gender (61% of the treatment group are female versus 45% of the comparison group) and by .35 units for age (students in the treatment group were 17.3 years old versus 17.1 years in the comparison group). For the college readiness sample, only one school was available from which to select comparison students. Our matching procedure more heavily weights parental educational attainment before the intervention began because that variable is more closely linked to our outcome. Because we prioritize this variable, and because the comparison pool was limited to a single school that also completed the *College Readiness Questionnaire*, we observe greater differences between treatment and comparison students on some other demographic variables. We include all available demographic variables in our analytic models.

[20] For quasi-experimental designs, the National Evaluation of i3 (NEi3) requires that researchers demonstrate that the treatment and comparison groups are balanced on a pre-determined variable. Per NEi3 guidance, a QED "must provide evidence that the groups being contrasted (e.g., the treatment and comparison groups) are equivalent at baseline on a pre-intervention measure of the outcome variable or on a variable or variables that are highly correlated with the outcome measure" (p. 14).

[21] To produce a standardized difference, we first produce model-based estimates of the difference between the treatment and comparison groups for each variable for each outcome domain for the primary (AVID-Core-Plus) and secondary (schoolwide) analyses. We then compute the pooled standard deviation for each variable. Finally, we produce a standardized difference of means by dividing the first term by the second.

[22] We use What Works Clearing House (WWC) standards to determine whether or not samples are balanced at baseline: a standardized mean difference of .05 or less satisfies baseline equivalence; for values between .05 and .25, WWC recommends adjusting statistical models to satisfy baseline equivalence; and values over .25 do not satisfy baseline equivalence.[22] Other researchers disagree regarding which value denotes a balanced sample; another recommendation is that a difference of less than .10 signifies a balanced sample. We include all observed covariates in our analytical models.

Results

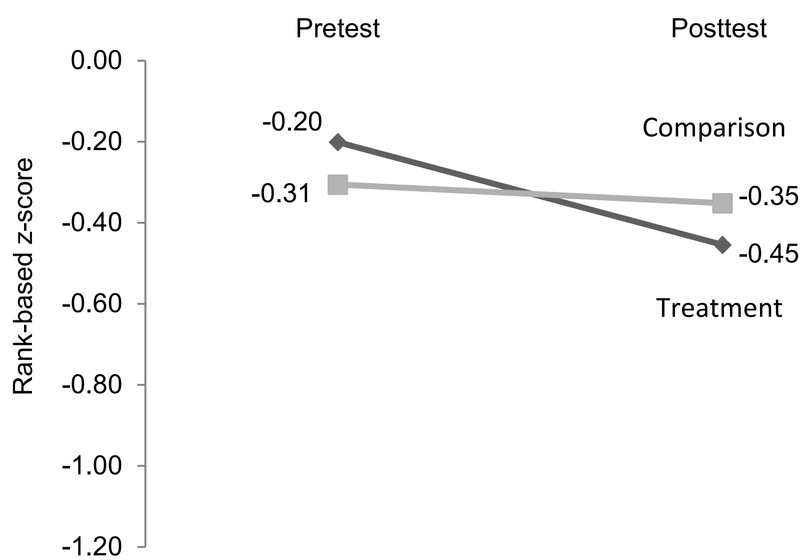
Academic Outcomes

The first two research questions of both the primary and secondary study investigate the observed impact of the AVID course and the schoolwide intervention together (primary study) and the schoolwide intervention alone (secondary study) on academic outcomes. Academic achievement is operationalized as the change in math and reading test scores from pre- to post-program. To estimate the impact of the interventions, we compare estimates of change in academic achievement over time for "treated" students to the change demonstrated by similar "comparison" students in other school districts in Arizona.

<u>Primary Study: Impact of AVID-Core-Plus on Academic Achievement</u>
Findings indicate that the AVID-Core-Plus intervention has no noticeable effect on academic achievement in math or reading.[23] Comparative estimates of pre- to post-program academic achievement produced by our statistical models are depicted in Figures 1 and 2. The pre to post trajectories of mean student achievement are slightly different for math and reading, but, in each case, the treatment and comparison groups are not different enough to provide evidence of program impact. In each case, the difference between the treatment and comparison groups' achievement change is not large enough to be statistically significant.

In analyzing the results for the primary study, we find that the AVID-Core-Plus intervention has no noticeable impact on the mathematics achievement of students exposed to the program. Figure 1 illustrates the change in relative performance on standardized mathematics tests for the treatment and comparison groups.

Figure 1. Relative Performance on Arizona State Math Test: Primary Study



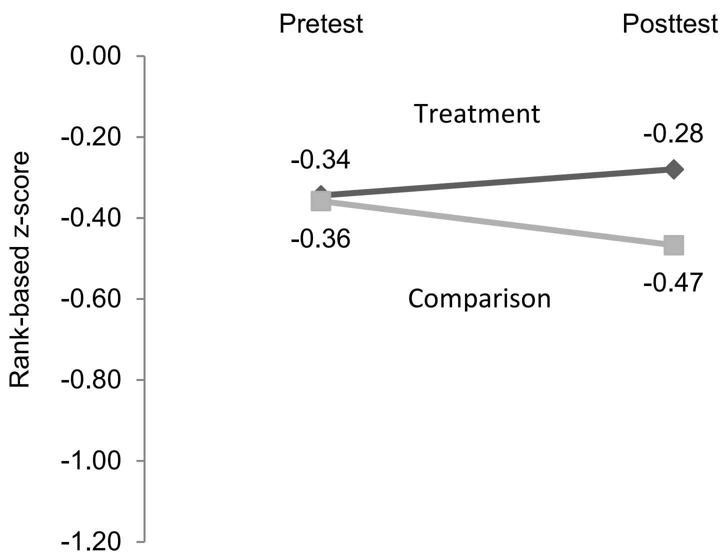
Note: Results are not statistically significant.

[23] Table G1 in Appendix G presents the results of our analytic model for academic outcomes in the primary study. The statistical model is specified in Appendix A.

The graphic illustrates the estimates of change in students' rank-based z-scores pre- and post-intervention.[24] Both treatment and comparison groups decline in their relative average ranking from pre- to post-intervention, but estimates indicate the mean value for the treatment group has decreased more (.25 standardized units) than the comparison group (.04 standardized units). The difference in change (-.21 standardized units) is the estimated "impact" of the program. The negative sign would seem to suggest a negative impact; however, the difference is small enough to fall within a range of uncertainty (i.e., it is not statistically significant). We conclude, therefore, that the AVID-Core-Plus intervention had no observable impact on math achievement.[25]

Results also indicate that the AVID-Core-Plus intervention has no noticeable impact on the reading achievement of program participants. Figure 2 illustrates the change in relative performance on standardized reading tests for the treatment and comparison samples.

Figure 2. Relative Performance on Arizona State Reading Test: Primary Study



Note: Results are not statistically significant.

Figure 2 shows that the treatment group improves its relative average ranking in reading test performance from pre- to post-intervention (.06 standardized units), and the comparison group decreases in its relative average ranking (-.11 standardized units). The difference in relative performance of the two groups (.17 standardized units) is in the desired direction in this analysis, but, again, it is small enough to fall within a range of uncertainty (i.e., it is not statistically significant). We conclude, therefore, that the AVID-Core-Plus intervention had no observable impact on reading achievement.

Secondary Study: Impact of Schoolwide Intervention on Academic Achievement

Findings indicate that the schoolwide intervention does not have a positive effect on academic achievement in math or reading. Students at BUSD who are exposed to the intervention actually demonstrate a greater decline in math and reading test performance over the course of the study than do comparison students selected from similar districts in Arizona.

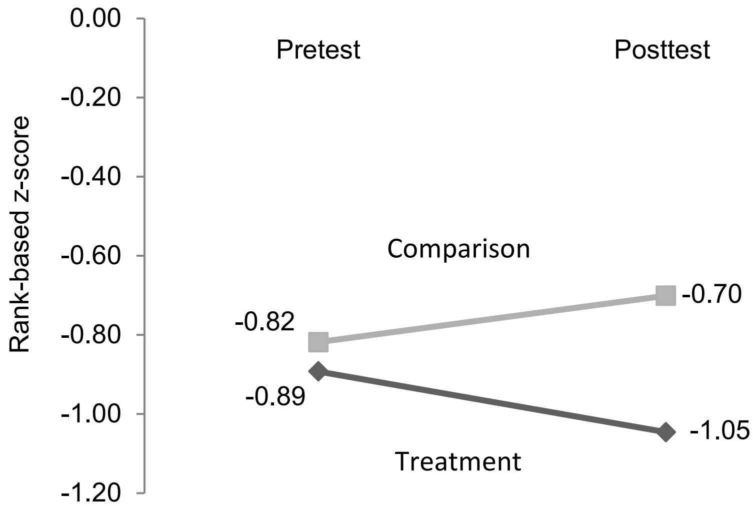
---

[24] The rank-based z-scores are a measure of students' performance on the standardized math test relative to the entire sample of students who took that test in Arizona in each school year.

[25] Table G1 in Appendix G presents the results of our analytic models for academic outcomes in the primary study.

Results from our statistical models indicate that the schoolwide intervention does not significantly improve relative standardized math test performance of the treatment group as compared to the comparison group. Instead, we find that the treatment group demonstrates a decline in math test performance that differs significantly from the slight gains demonstrated by students in the comparison group.[26] Figure 3 plots the relative change in math test performance for the treatment and comparison group from pre- to post-program.

Figure 3. Relative Performance on Arizona State Math Test: Secondary Study



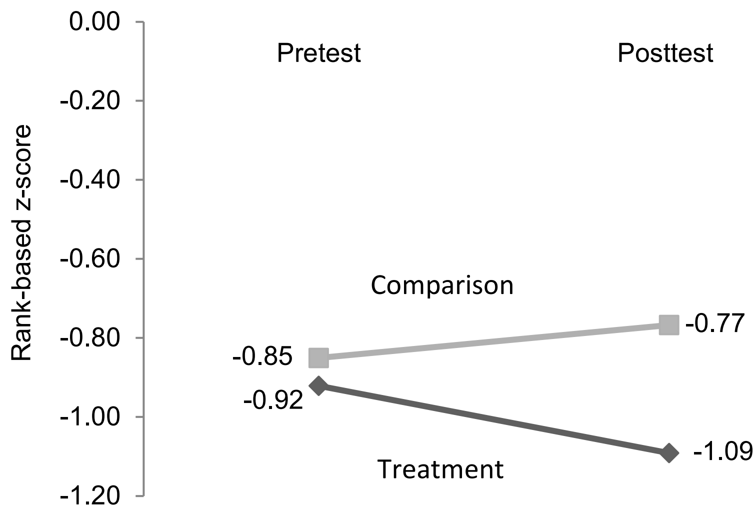
Note: Results are statistically significant (*p* < .05).

Pre- to post-intervention estimates of impact on relative math test performance produced by the statistical model are illustrated in Figure 3. Figure 3 shows that the treatment group decreases in its relative average ranking from pre- to post-intervention (-.16 standardized units), and the comparison group increases in its relative average ranking (.12 standardized units). The difference in change between the two groups (-.27 standardized units) is statistically significant ($p < .05$).

Similarly, results indicate that the schoolwide intervention does not positively impact the reading test performance of students. Contrary to our expectations, we find that relative test scores of students in the treatment group decrease relative to scores of students in the comparison group.[27] Figure 3 plots the relative change in reading test performance for the treatment and comparison group from pre- to post-program.

[26] Table G2 in Appendix G presents the results of our analytic models for academic outcomes in the secondary study.
[27] See Table G2 in Appendix G.

Figure 4. Relative Performance on Arizona State Reading Test: Secondary Study



Note: Results are statistically significant (*p* < .001).

Pre- to post-intervention estimates of reading test performance produced by the statistical model are illustrated in Figure 4, which shows that the treatment group decreases in its relative average ranking from pre- to post-intervention (-.17 standardized units), and the comparison group increases in its relative average ranking (.08 standardized units). The difference between the two groups (-.25 standardized units) is statistically significant ($p < .001$).

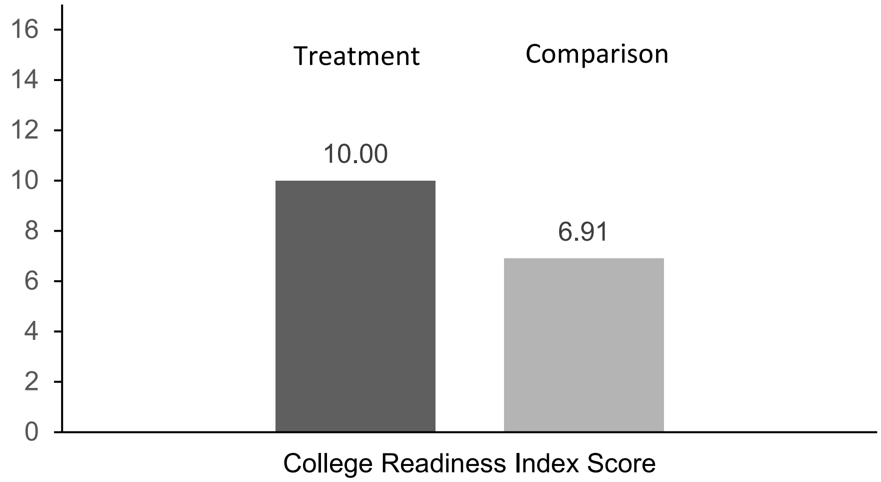
College Readiness

The final research question for both the primary and secondary studies investigates the impact of AVID plus the schoolwide intervention (primary study) or the schoolwide intervention (secondary study) on students' college readiness at the end of the intervention. We use the CRI as a measure of students' self-reported college readiness. The CRI is a 17-item index that measures students' academic preparedness, college knowledge, and college admissions preparedness; scores are on a scale of 0 to 17. The index quantifies the extent to which a student self-reports that he or she is prepared for enrolling in college. College readiness is assessed at posttest only, in the final year of the intervention (2014-15 school year).

<u>Primary Study: Impact of AVID-Core-Plus on College Readiness</u>
Findings indicate that the AVID-Core-Plus intervention improves students' college readiness. Students who are exposed to AVID-Core-Plus have significantly higher levels of college readiness post-program than do similar students not exposed to the program, as measured by CRI scores. Figure 5 illustrates the difference in college readiness between the two groups. AVID-Core-Plus students have a mean CRI score of 10 (on the 17-point scale), compared to a mean score of 6.9 among students in the comparison group.[28] The difference in CRI scores between the two groups (3.1 points) is the estimated impact of the program and is highly statistically significant ($p < .001$).

[28] Table G3 in Appendix G describes our full analytic model predicting college readiness.

Figure 5. Mean College Readiness Index Scores for Treatment and Comparison Students, at Posttest: Primary Study

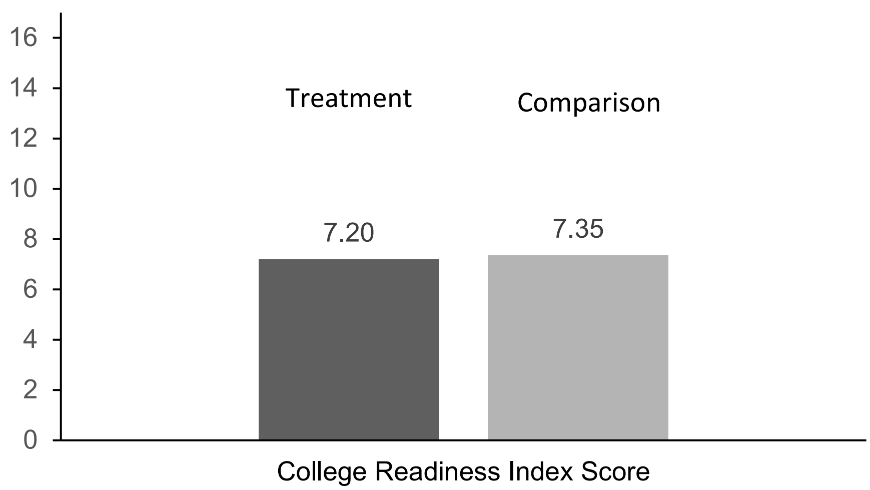


Note: Results are statistically significant ($p < .001$).

Secondary Study: Impact of Schoolwide Intervention on College Readiness

Results indicate that the schoolwide intervention had no noticeable effect on the college readiness of study participants. CRI scores for the treatment and comparison groups for this contrast are quite similar, as illustrated by Figure 6. Students in the schoolwide treatment group have a mean CRI score of 7.20 (on the 17-point scale), compared to a mean score of 7.35 among students in the comparison group. The difference in CRI score between the two groups (-.15 points) is the estimated impact of the program; the statistical model also indicates, however, that this difference is statistically insignificant.

Figure 6. Mean College Readiness Index Scores for Treatment and Comparison Students, at Posttest: Secondary Study



Note: Results are not statistically significant.

Conclusion and Study Limitations

The primary study in this impact evaluation assesses the effect of the AVID-Core-Plus intervention on student academic outcomes and college readiness. The secondary study assesses the impact of the schoolwide intervention on the same outcomes. Using non-experimental statistical techniques, we estimate impacts by contrasting outcomes for students exposed to these interventions with outcomes of students not exposed to the interventions. The results of our primary study suggest that the AVID-Core-Plus intervention does not impact students' achievement in math or reading, but that it does appear to improve the college readiness of students in the 11th and 12th grades. In our secondary analyses, we find that the schoolwide intervention does not improve either academic or college readiness outcomes.

The evaluation team believes that the design and analytic methods employed here are the most rigorous available given the constraints initially imposed. Nevertheless, because these results are derived from non-experimental methods, they are subject to a number of limitations. First, as with all non-experimental methods, the validity of the findings presented in this report rest on the assertion that, conditional on statistical adjustments, the treatment and comparison groups are equivalent. This contention is partially verifiable by examining the distribution of observed variables at baseline for the contrasted groups (i.e., treatment and comparison), which, in all analytic samples, were quite well-balanced. Ultimately, however, we cannot be sure that the two groups are, in fact, similar in terms of the many characteristics that go unobserved or are practically unobservable. If there are differences between the two groups that we cannot account for in our statistical models, these differences may contribute to or have an effect on the study outcomes, introducing bias. It is worth emphasizing that, because of design constraints, important attributes that differentiate our two groups and may influence academic and college readiness outcomes have gone unobserved.

An additional design constraint that further qualifies the results presented here is the unavoidable overlap of the implementing school district and the interventions being evaluated. Students who are exposed to the treatment interventions (AVID-Core-Plus and schoolwide) come from a single district, and comparison students come from other districts. This perfect alignment of district and intervention means that we are unable to effectively distinguish district effects from intervention effects. While the DID analytical approach employed in the analysis of academic outcomes helps to control for baseline differences between the treatment and comparison groups, which may include district-level contextual factors that influence academic performance, it assumes a parallel growth or decline in terms of those outcomes at baseline. This means that any baseline district-level differences that could motivate differential trends in the outcomes (other than the interventions) could be influencing (at least, in part) the results reported here.

.

Appendix A: Methods

Overview

This impact study investigates observed effects of two separate interventions on student-level math and reading (academic) achievement and college readiness outcomes. For each of these outcomes, we assess two contrasts: 1) an exploratory contrast that compares students exposed to a schoolwide intervention (but not AVID core coursework) with individually matched students who attend a similar district; and 2) a confirmatory contrast that compares students exposed to the schoolwide intervention plus an AVID-core program (AVID-Core-Plus) with individually matched students who attend a similar district.

This appendix provides details on sample selection, data collection, and analytic methods for this study. This study is a Quasi Experimental Design (QED) whose purpose is to isolate the impact of the intervention on student-level academic achievement and college readiness outcomes. For the achievement outcomes, the treatment effect is estimated by way of a difference-in-differences (DID) method that is common in program evaluation and education research (e.g., Angrist and Pischke, 2009; Schlotter et al., 2010; Antonakis et al., 2011; Murnane and Willett, 2011).[29] College readiness outcomes are assessed at post-test only. We use multi-level empirical models to produce impact estimates, controlling for relevant covariates at the individual- and school-level.

Consistent with the WWC and NEi3 definitions of a QED, we establish baseline equivalence in our treatment and comparison groups for each outcome/contrast sample.[30] Analytic samples have been constructed by matching students in the treatment group with an equivalent group of comparison students who attend other school districts in Arizona with similar important characteristics (e.g., proportion Native American, rural and remote, economic resources, state academic performance rating), but are not receiving the intervention. Students in each grade who participate in the treatment conditions (core and schoolwide) and have pre- and post-intervention outcome data are matched individually by way of matching procedures that take into account characteristics that are associated with the selection process and outcome measure. These characteristics are measured pre-intervention and include baseline achievement on the outcome measure, if available, as well as other important student-level characteristics that are thought to predict selection into the intervention in addition to the outcome itself. We establish baseline equivalency in line with NEi3 and WWC guidelines.

Since the Arizona Department of Education changed the standardized tests for students in the final year of our study, outcomes are operationalized as rank-based z-score transformations of student scaled scores on standardized state reading and mathematics tests required by the state of Arizona. College readiness outcomes are operationalized as the student's self-reported score on the college readiness index (CRI) which is included as part of a questionnaire that was administered in the spring semester of the final year of the intervention.

[29] Angrist, J. D. and J. S. Pischke (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton Univ Press. Schlotter, M., Schwerdt, G. and Woessmann, L. (2010). Econometric Methods for Causal Evaluation of Education Policies and Practices: A Non-Technical Guide, Institute for the Study of Labor (IZA) Discussion Paper No. 3478. Available online: http://ftp.iza.org/dp4725.pdf. Antonakis, J., Bendahan, S. Jacquart, P., and Lalive, R. (2010). On Making Causal Claims: A Review and Recommendations, The Leadership Quarterly, 21, 1086-1120. Murnane, R.J. and Willett, J.B. (2011) Methods Matter: Improving Causal Inference in Educational and Social Science Research. New York: Oxford University Press.

[30] A Quasi Experimental design is defined in the What Works Clearinghouse Procedures and Standards Handbook as a design that "compares outcomes for students, classrooms, or schools who have access to the intervention with those who did [sic] not but are similar on observable characteristics." (p. 11, WWC, v. 3.0). Similarly, the NEi3 defines the QED as a study that "compares outcomes for intervention participants with outcomes for a comparison group chosen through methods other than randomization." (p.14).

Sample Selection

Control (or Comparison) Conditions

Students at comparison schools receive any instruction that is offered by the school/district (i.e., business as usual). In this case "business as usual" reflects the typical curriculum and programming that may include some college readiness activities. AVID makes available on its website a list of schools and districts implementing AVID in Arizona; we used the list to confirm that none of the selected comparison districts were implementing AVID during the grant period.

Sample Identification, Selection, and Assignment

The study examines academic achievement and college readiness outcome effects of an intervention that was administered in a single district in the state of Arizona. Two treatment conditions were offered at the single middle school (BMS) and the single high school (BHS) in that district. Students in those schools were selected into one of two conditions: 1) a "schoolwide" condition that constitutes the intervention for the exploratory contrast for all outcomes in the study; and 2) an AVID-Core-Plus schoolwide condition that constitutes the confirmatory contrast for all outcomes in the study.[31] If a student was enrolled in the AVID-Core-Plus condition for at least one year, s/he was selected into the AVID-Core-Plus treatment condition. If a student attended one of the treatment schools (BMS and BHS) during the intervention period, but was not selected into the AVID-Core-Plus condition, s/he was "selected into" the schoolwide condition.[32]

For academic achievement outcomes, the availability of student test scores (Arizona only requires testing in grades 6-8, and 10) mean that outcomes are not assessed for students in grades 9, 11, and 12. For college readiness outcomes, only students in grades 11 and 12 in the final year of the intervention period are included in the analytic sample.

The study examines AVID core-plus and schoolwide intervention impacts on academic (math and reading) and college readiness domains with six contrasts (i.e., 3 domains x 2 contrasts = 6 outcomes). For each of these six contrast conditions, we match treatment with comparison samples by matching students in the treatment sample with individuals who are similar in terms of theoretically relevant baseline characteristics and who come from districts that are similar to the treatment district. The comparison districts are selected differently for academic and college readiness outcomes, from a pool of five districts that have been identified as being equivalent to the treatment district.

For academic outcomes, students who attend (non-alternative) middle or high schools in each of the five districts during the intervention period are included in the matching pool and thus may potentially be members of one or both of the comparison samples (for both exploratory and confirmatory contrasts). Individual-level matching procedures identify the final comparison samples.[33] The primary research question for this pair of contrasts concerns the impact of the intervention on the relative placement of students in terms of test scores from pre- to post-intervention.

For college readiness outcomes, only students who attend high school in one anonymous comparison district from the selected five districts are included in the matching pool and thus may potentially be

[31] Although the schoolwide intervention was provided to all students in the treatment schools, including the students in the AVID-Core-Plus condition, only those students who received only the schoolwide intervention constitute the schoolwide treatment sample. Therefore, speaking in terms of study samples, those students who are selected – and elect to enroll in the AVID-Core-Plus condition – are selected into that condition. Students who are not, are "selected into" the schoolwide sample.
[32] Both these statements assume that the student provides sufficient data to be selected into the analytic sample. See below for clarification.
[33] Again, this statement presumes these comparison students have sufficient outcome data.

members of one or both of the comparison samples (for both exploratory and confirmatory contrasts). Individual-level matching procedures identify the final comparison samples. The primary research question for this pair of contrasts concerns the impact on the level of student-reported college readiness after the intervention was operational at BUSD schools for three years.

Identification/Selection of Study Districts

We have conducted a student-level analysis; however, prior to matching students, we first pre-screened for districts that are comparable to the treatment district on key characteristics (Somers, Zhu, Jacob, & Bloom, 2013).[34] In other words, we identify a matching/sampling frame from specific districts in Arizona that are similar to the treatment district. Out of an initial seven districts that were identified as similar according to five key characteristics (see Table 1), five were eligible for pre-selection. For college readiness outcomes, we obtained an MOU with one of these districts and administered our college readiness questionnaire at the single secondary school in the district.

Selection of Comparison Districts for Academic Achievement Outcomes

Prior to individual-level matching, we first pre-screened the matching pool by selecting a set of school districts inside Arizona that are comparable on relevant district-level characteristics.[35] Research indicates that observational methods are much more successful at reducing the effects of selection bias when the comparison sites are as similar as possible – optimally geographically proximate to the treatment site (Cook, Shadish, & Wong, 2008).[36]

Baseline (i.e., pre-intervention) characteristics that were considered are: (1) proportion of the school district that is Native American; (2) urbanicity of the district; (3) past academic performance of the school district; (4) population of the district; and (5) economic indicators. A preliminary analysis of the school districts in Arizona indicated that districts are either predominantly Native American or not. When districts are sorted by this proportional distribution, an evident cut-point occurs in districts that are around 90% Native American. In Table A1, we present every district with 90% Native American student population or greater in the state of Arizona. For each of these districts, we also provide statistics on the additional district-level matching characteristics.

Based on the criteria considered, five school districts were selected as district-level comparable to BUSD. Each of the districts is predominantly Native American, rural and remote/distant, and rated similarly by the Arizona Department of Education in terms of their academic performance in 2011. Additionally, the districts appear to be alike in terms of median household income and size (population). Eight districts that are predominantly Native American were not selected because they are substantially different to BUSD on one or more of the other characteristics, or because they do not include a non-alternative high school.

[34] Somers, M. A., Zhu, P., Jacob, R., & Bloom, H. (2013). The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation. *MDRC*.

[35] This is consistent with the guidance offered by Song and Herman (2010). Song, M. and Herman, R. (2010). Critical Issues and Common Pitfalls in Designing and Conducting Impact Studies in Education: Lessons Learned from the What Works Clearinghouse (Phase I). Educational Evaluation and Policy Analysis, 32, 351-371.

[36] Cook, T. D., Shadish, W. R., and Wong, V. C. (2008). Three conditions under which experiment and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, 27(4), 724–750.

Table A1. Baseline Characteristics of Selected and Not Selected School Districts with Primarily Native American Population in Arizona[37]

	District Population [38]	Percent Native American	AZDOE Letter Grade[39]	Urban Code and Description[40]		Median Income
Baboquivari Unified Dist.	**5,896**	**96.8%**	**D**	**43**	**Rural/Remote**	**$24,496**
Selected Comparison Districts:						
A	6,028	89.3%	D	43	Rural/Remote	21,431
B	5,159	91.9%	D	43	Rural/Remote	25,704
C	9,492	95.1%	D	42	Rural/Distant	22,343
D	5,425	96.7%	D	43	Rural/Remote	21,399
E	6,485	93.1%	D	43	Rural/Remote	29,561
Districts Not Selected:						
F[41]	9,287	97.4%	D	43	Rural/Remote	32,171
G[42]	7,346	97.0%	D	42	Rural/Distant	31,544
H	18,033*	96.1%	D	42	Rural/Distant	20,161
I	8,653	96.1%	C*	43	Rural/Remote	24,754
J	7,767	99.4%	C*	43	Rural/Remote	21,343
K	15,227	92.5%	D	43	Rural/Remote	40,014*
L	11,187	91.2%	D	33	Town/Remote*	30,940
M	12,042	97.7%	D	33	Town/Remote*	33,525

*Indicates criterion on which the district was excluded.

Selection of Comparison District for College Readiness Outcomes
One comparison district agreed to administer the annual *College Readiness Questionnaire*. We agreed to keep the district anonymous for the purposes of the study.

<u>Identification/Selection of Students for the Impact Evaluation</u>
In this section, we outline the procedures for identifying what students are included in the treatment samples and for identifying students who were then selected as part of the matched comparison samples.

Students are operationally defined as participating in the treatment arms if they are enrolled in BHS or BMS and have complete outcome data. They are assigned to the exploratory or confirmatory contrast on the basis of enrollment in the AVID-Core-Plus program. They are included in the specific outcome samples on the basis of criteria specified below.

[37] Letters are used instead of district names to help maintain the anonymity of the comparison districts.
[38] Source for district-level statistics on population, Percent AI, and Median Income: U.S. Census Bureau; American Community Survey, 2006-2010. Obtained from National Center for Education Statistics. http://nces.ed.gov/ccd/CCDLocaleCodeDistrict.asp; March 2, 2013.
[39] Source for data for this column: Arizona Department of Education Research and Evaluation, Accountability Web Page. http://www.azed.gov/research-evaluation/a-f-accountability/; March 1, 2013.
[40] Source for data in this column: National Center for Education Statistics, Local Education Agency (School District) Locale Code Files, http://nces.ed.gov/ccd/CCDLocaleCodeDistrict.asp; March 1, 2013.
[41] District F met criteria for baseline characteristics but was not selected because it only includes an alternative high school.
[42] District G met criteria for baseline characteristics but was not selected because the district does not include a high school.

Separate comparison samples are identified for each of the six outcome/contrast combinations according to the procedures outlined below. Students were eligible for selection into a comparison sample if administrative tests (for academic outcomes) or questionnaire data indicate they attended school in a pre-selected district (see above) during the intervention period for each relevant cohort and have sufficient data. Students in the treatment condition were then matched (based on relevant background characteristics) to a student in the specific outcome/contrast treatment sample. The matching procedures seek to maximize the equivalence of the treatment and comparison groups in terms of key observed variables by selecting only those students who most closely approximate the treatment students in terms of several key variables (specified later in this section).

General Criteria
The first requirement for inclusion into the study – equally relevant for treatment and comparison samples – is sufficient outcome data. This is actualized differently for each of the outcome analyses, but the principle remains the same. Since we do not impute outcome data in our study, students cannot be included in the empirical estimation of intervention effects if they do not have sufficient outcome data. We detail the criteria by outcome domain below:

- **Academic Outcomes**. To be eligible for inclusion into the analytic sample for academic outcomes, students must have data from the state-required standards-based test for both pre- and post-intervention. The pre and posttests must be linkable (by student ID) and in the same subject. A student may therefore be eligible to be in the analytic sample if s/he has a reading or a math test score in the identified pretest occasion for his/her cohort (see below) and a reading or math test score at the identified post-intervention occasion (see below).[43] In Arizona, students are required to take standardized tests in grades 3 through 8 and again in grade 10. These tests are high-stakes and high participation (May, et al., 2005). Students in grades 6 through 12 were exposed to either the AVID-Core-Plus or schoolwide intervention in the study period, but only students who completed a posttest in grades 6, 7, 8, or 10 at the specified pre and post- intervention administrations are included in the respective analytical sample.[44]

- **College Readiness Outcomes.** To be eligible for inclusion into the sample for the analysis of college readiness outcomes, a student must have been in grades 11 or 12 and responded to all the items in the college readiness index and answered the items on parent education during the administration of the *College Readiness Questionnaire* (CRQ) in the final year of program implementation (spring 2015).

Identification of Treatment Students
In this section we describe our inclusion, exclusion, and timing criteria for selecting students into the treatment arm of the study for both outcomes. Tables A2 and A3 below outline cohort progression through the AVID-Core-Plus and schoolwide interventions. Colors are used to illustrate each of the cohorts as they move through the intervention. Additionally, a letter/number indicates the intervention year (a, b, c) and grade where treatment cohorts were first exposed to the treatment condition.[45] We also indicate the start of the intervention with an outline. Each of the cohorts includes all students who were enrolled in either treatment contrast (in AVID-Core-Plus or schoolwide) by the final possible intervention observation point.

---

[43] As noted in the model specification section, these procedures assume that we are using a single pre- and post-test observation in the modeling of program impacts on academic achievement.
[44] In 2015, AZDOE changed from the AIMS standardized test to the AZMerit test. As part of this change, 10th grade students are no longer required to take an end of grade math test, beginning in the spring of 2015. Students who were in 10th grade in spring 2015 (cohort A8) were not required to take an end of grade math test and were excluded from analysis.
[45] Note, however, that a student may be selected into either treatment contrast after the initial cohort exposure year.

Table A2. Progression of Cohorts through Core-Plus Intervention

	2011-2012	2012-13	2013-14	2014-15
Grade 5				
Grade 6			B6	C6
Grade 7		A7	B7*	
Grade 8		A8		
Grade 9		A9		
Grade 10		A10		
Grade 11				
Grade 12				

Table A3. Progression of Cohorts through Schoolwide Intervention

	2011-2012	2012-13	2013-14	2014-15
Grade 4				
Grade 5				
Grade 6		*A6	B6	C6
Grade 7		A7		
Grade 8		A8		
Grade 9		A9		
Grade 10		A10		
Grade 11		*A11		
Grade 12		*A12		

***Note:** In the schoolwide intervention, cohort A6 has different first year of intervention than in core-plus (B7 in core-plus). Cohorts A11, A12 are not included in the core-plus intervention.

Tables A2 and A3 illustrate the seven and nine cohorts who have the potential to be exposed to the core and schoolwide interventions, respectively.

Academic outcomes sample. Our research questions are concerned with determining if and to what extent the intervention impacted the relative improvement of students' academic achievement. As illustrated in the tables above, AVID-Core-Plus and schoolwide interventions were implemented in BMS and BHS in grades 6 through 12 over the course of the study period (fall 2012 to spring 2015). Ideally, we would be able to include students in each of these grades in the analysis and measure them after similar exposure. However, test data are not available for all grades, and the AVID-Core-Plus intervention delayed implementation with 6th grade students by one year. The AVID-Core-Plus intervention was operative in grades 7 through 10 in the first year (2012-13), grades 6 through 11 in the second year (2013-14), and all grades from 6 through 12 in the final year (2014-15).

Since pre and/or post testing data are not available for every one of these cohorts, some are not eligible for inclusion into the study (see Tables A4 and A5 below). For the core intervention analysis, cohort A8 is excluded from the analytic sample because end-of-grade standardized post-test data are not available.[46] For the schoolwide intervention analysis, cohorts A11 and A12 are excluded in addition to A8 because posttest data are not available for these students. In all, six cohorts are included in the analytic samples for both the exploratory and confirmatory contrasts.

College readiness outcome sample. Our research question is concerned with determining whether or not students exposed to the intervention have higher levels of self-reported college readiness than similar students not exposed to the intervention. As is shown in Tables A2 and A3, by the final year of the study period, both the AVID-Core-Plus and schoolwide interventions are being implemented in BMS and BHS in all grades. We are not able to match students and conduct an analysis that is longitudinal in structure. The study is a post-only comparison group design, and only those participants in grades 11 and 12 in the final year of the intervention study period – cohorts A9 and A10 – provide outcome data.

Matching Procedures for Selecting Comparison Samples

Although the intent is to make causal inferences about the impact of the intervention, ignoring treatment assignment is not credible in this case. First, we have not conducted an RCT; second, treatment assignment is confounded with our intervention district. Nevertheless, we have endeavored to balance our analytic samples in a way that approximates the effects of randomization – at least on our observed key variables. The selected procedure involves selecting a comparison sample that is balanced in terms of relevant baseline characteristics that may influence selection into treatment and outcomes. It also involves selecting individuals who meet the same basic data requirements as the treatment group (outlined above). First we provide details on the three steps we take to match treatment participants to comparison participants in a way that generates balanced samples (at least in terms of measured variables).

Step 1: Narrow Analytic Sample to Comparable School Districts. This is detailed above.

Step 2: Identify Variables to Include in Match. Guidance from propensity score matching literature encourages researchers to include all variables that are theoretically expected to be related to the treatment assignment and the outcome in the matching procedure (e.g., Guo & Fraser, 2010; Stuart, 2010). On balance, one is encouraged to be more inclusive in selecting matching variables (Rubin & Thomas, 1996; Hill et al., 2004; Stuart & Rubin, 2007; Stuart, 2010).[47] However, findings and guidance are not unanimous on this matter, and the more recent literature suggests that, when faced with restrictions, researchers should give priority to covariates that are related to the outcome rather than treatment assignment (Guo & Fraser, 2010; Stuart, 2010). At a minimum, where available, Song and Herman (2010) advise that "matching should be done on a pre-intervention measure (pretest) of the outcome or a close proxy measure for the pretest" (p. 355).[48] We follow their guidance here. For matching our academic outcomes samples, we include a pre-intervention measure of the outcomes (math and/or reading AIMS pre-test scores) as well as all other variables that we expect are highly correlated with the outcome. For matching our college readiness outcomes samples, we include a measure that is

[46] In 2015, AZDOE changed from the AIMS standardized test to the AZMerit test. As part of this change, 10th grade students are no longer required to take an end of grade math test, beginning with the spring of 2015.

[47] Rubin, D. B., & Thomas, N. (1996). Matching Using Estimated Propensity Scores, Relating Theory to Practice. *Biometrics, 52,* 249–264. Hill, J., Reiter, J., & Zanutto, E. (2004). A Comparison of Experimental and Observational Data Analyses. In A. Gelman & X.L. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective* (pp. 44–56). New York: John Wiley. Stuart, E.A. and Rubin, D.B. (2007). Best Practices in Quasi-Experimental Designs: Matching Methods for Causal Inference. Chapter 11 (pp. 155-176) in *Best Practices in Quantitative Social Science*. J. Osborne (Ed.). Thousand Oaks, CA: Sage Publications.

[48] Similarly, Stuart (2010) writes that matching should be done on the most prognostic variable. The NEi3 (2012), advises that in assessing baseline equivalence, the researcher should include one "variable that is likely to be highly correlated with the outcome variable" (p. 16). The only variables that one should not include are those that may have been affected by the treatment of interest, as this can lead to bias in the estimated treatment effect (Imbens, 2004). Song, M., & Herman, R. (2010). Critical Issues and Common Pitfalls in Designing and Conducting Impact Studies in Education. *Educational Evaluation and Policy Analysis*, 32, 351–371. Imbens, G., 2004, Nonparametric Estimation of Average Treatment Effects under Exogeneity:
A Review, *Review of Economics and Statistics*. 86, 4-30.

theoretically highly correlated with the outcome (parents' education) as well as all other available variables that we expect are highly correlated with the outcome. With all this in mind, we selected the following variables for inclusion in the creation of matching scores. The asterisk in each outcome domain indicates the variable that we identified in the analysis plan to demonstrate baseline equivalence, as detailed below, in the contrast tool and in the baseline balance testing section of this plan, and as stipulated by NEi3 *Analysis and Reporting Plan*.

Academic Achievement Outcome Matching Variables

- Gender
- Native American indicator
- English language learner status indicator
- Special education status indicator
- Free and reduced-price lunch status indicator
- Pre-intervention test (pre-test) score(s)*

College Readiness Outcome Matching Variables

- Native American indicator (Race/Ethnicity)
- Gender/sex
- Age
- Grade [49]
- Living with both biological parents
- Household resources (Measure of SES)[50]
- Parents' education (Measure of SES)*

Step 3: Distance Metric and Matching Method. The distance metric we selected for matching is Euclidean distance with propensity score calipers. Euclidean distance matching is recommended for small samples and a limited number of covariates. Further, since there is a single covariate that is more predictive of the outcome variable than the others (the pretest), Euclidean distance matching is ideal because it can incorporate this primacy in the distance calculation by weighting that covariate more than the others. The use of calipers and caliper settings were based on recommendations in Rosenbaum and Rubin (1985).[51]

The Euclidean distance captures the straight-line distance between two points in an *n*-dimensional space, with *n* being the number of variables and the two points being observations of those variables. For the core academic outcomes, we calculated the Euclidean distance as follows:

$$\sqrt{(T_{pt1} - C_{pt1})^2 - \left(T_g - C_g\right)^2 - (T_{ELL} - C_{ELL})^2} - (T_{SP} - C_{SP})^2 - (T_{FR} - C_{FR})^2 - (T_{NA} - C_{NA})^2$$

For the schoolwide academic outcomes, we calculated the Euclidean distance as follows:

$$\sqrt{(T_{pt1} - C_{pt1})^2 - \left(T_{pt2} - C_{pt2}\right)^2 - \left(T_g - C_g\right)^2 - (T_{ELL} - C_{ELL})^2} - (T_{SP} - C_{SP})^2 - (T_{FR} - C_{FR})^2 - (T_{NA} - C_{NA})^2$$

[49] Grade was not used in matching process since students were matched by grade.
[50] See College Readiness Survey Q.9 in Appendix E.
[51] Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician, 39*(1), 33–38.

For the core and schoolwide college readiness outcomes, we calculated the Euclidean distance as follows:

$$\sqrt{(T_{pe1} - C_{pe1})^2 - \left(T_g - C_g\right)^2 - (T_{HH} - C_{HH})^2 - (T_{LP} - C_{LP})^2 - (T_A - C_A)^2 - (T_{NA} - C_{NA})^2}$$

Where *T* refers to an individual in the treatment group; *C* refers to an individual in the comparison group; *pt*1 is the pre-test score in the same subject as the outcome; *pt*2 is the pre-test score in a different subject as the outcome (i.e., English if outcome is math and math if outcome is English); *pe* is a parental education indicator; *g* is gender; *ELL* is English language learner status; *SP* is special education status; *FR* is free/reduced lunch status; *HH* is a household resources indicator; *A* is an age indicator; *LP* is a lives-with-both-parents indicator; and *NA* is a Native American indicator. In the first equation, the first pre-test is weighted more heavily (two times) than the other variables. In the second equation, the second pre-test has a weighting that is equal to the other background characteristics. In the third equation, the parental education is weighted more heavily (two times) than the other background characteristics.

Matching is conducted for each outcome/contrast within each of the six cohorts (see Tables 4 and 5 below). Within each contrast/cohort set, we calculate a distance for all possible treatment-comparison pairs and then select comparison cases on the basis of a 1:1 closest match, without replacement, for each treatment/comparison pair that minimizes the Euclidean distance. Specifically, matching is done by identifying the closest comparison match within each cohort for each treatment case, removing that pair, and then matching the next closest treatment-comparison pair. This routine continues until all eligible treatment cases have been matched with a comparison case within each cohort.

Since we have decided to use propensity score calipers, only those matched cases whose (absolute) difference in propensity scores do not exceed .25 of a standard deviation are eligible to be included in the analytic sample. Propensity scores using the same set of covariates were estimated with logistic regression. For academic outcomes, propensity scores were estimated with the following variables: gender, English language learner status, special education status, free and reduced lunch status, and pre-intervention test score for same subject as the outcome. Propensity scores for the college readiness samples were calculated with the following variables: gender, age, grade, lives-with-both-parents indicator, household resources indicator, and parent education indicator.

All variables used in the matching equations were either collected prior to the intervention (e.g., pretest scores) or reflect characteristics that were not expected to change as a result of the intervention. For clarification on when baseline data were gathered for matching purposes, see Tables 4 and 5, below.

Academic Achievement Outcome Samples

Any student who is identified as being in the AVID-Core-Plus treatment (confirmatory contrast) for at least one year and who has sufficient data is selected as a study treatment participant for the core contrast (reading and/or math). Whereas not every student who is a member of one of the six identified cohorts will have sufficient testing data, only members of those cohorts have the potential to provide sufficient pre- and posttest data (see Tables 4 and 5 for clarification). Any student who is selected as a treatment student in the core-plus sample is not eligible to be selected into the schoolwide treatment group. The schoolwide treatment group is, therefore, any student who attended BMS or BHS, was not a core participant prior to their posttest, is in one of the identified six cohorts, and has sufficient data.

Comparison students are considered eligible for matching purposes if they attended one of the identified schools (i.e., not alternative) in the five eligible districts during the treatment period and testing data indicate that they are in one of the six relevant cohorts (for example, cohort A9, whose members were in the eighth grade in 2012, the tenth grade in 2014, and have testing data for both years), and they have sufficient data. Comparison students may be in both confirmatory and exploratory samples, but treatment students cannot be in both.

<u>College Readiness Outcome Samples</u>

We were not able to gather student-level data prior to the intervention, nor are we able to reliably link data obtained in the questionnaire with other student-level data (i.e., test data). As such, we match individuals based on data gathered in the *College Readiness Questionnaire* instrument that was administered in the final year of the intervention (2014-15). Any student who completed the questionnaire and indicated that he/she has participated in the AVID-Core-Plus program is considered a member of the core-plus treatment group. All other respondents who completed the questionnaire at BHS are considered schoolwide participants. Students who completed the questionnaire at the comparison high school are considered eligible for matching for either (and both) the confirmatory and exploratory contrasts.

<u>Academic Achievement Outcome Measurement</u>

Since the level of measurement and assignment into the study is at the individual level, a student may enter the AVID-Core-Plus (or leave the schoolwide) sample at any point in the three-year intervention period and not necessarily be exposed for the full two or three possible years of exposure for their cohort. Students are assigned to the **confirmatory contrast** if, on the basis of student enrollment data, they were enrolled for at least one year in the AVID-Core-Plus intervention and they have sufficient outcome data. Students are assigned to the **exploratory contrast** if the administrative data indicate that they were not enrolled in the AVID-Core-Plus intervention prior to their posttest and have sufficient outcome data.

Students in the academic outcomes samples will experience the intervention in one of six cohorts, with possible exposures ranging from one to two years in the confirmatory contrast and one to three years in the exploratory contrast. Tables 4 and 5 below specify the grade/year for pre-intervention (pre) measures, post-intervention (post) observations, and the possible years of exposure for each cohort.

Table A4. Confirmatory Contrast: Multi Year Intervention Measurement and Exposures by Cohort

	2010-2011	**2011-2012**	**2012-13**	**2013-14**	**2014-15**
Grade 4			pre C6		
Grade 5		pre B7	pre B6		
Grade 6		pre A7			post C6
Grade 7					post B6
Grade 8	pre A10	pre A9		post A7	post B7
Grade 9					
Grade 10			post A10	post A9	
Grade 11					
Grade 12					

Note. Maximum years of exposure by cohort are as follows: C6 = 1 year, B6 = 2 years, B7 = 2 years, A7 = 2 years, A9 = 2 years, A8 = 3 years, A10 = 1 year.

Table A5. Exploratory Contrast: Multi Year Intervention Measurement and Exposures by Cohort

	2010-2011	2011-2012	2012-13	2013-14	2014-15
Grade 4			pre C6		
Grade 5		Pre A6	pre B6		
Grade 6		pre A7			post C6
Grade 7					post B6
Grade 8	pre A10	pre A9		post A7	post A6
Grade 9					
Grade 10			post A10	post A9	
Grade 11					
Grade 12					

Note. Maximum years of exposure by cohort are as follows: C6 = 1 year, B6 = 2 years, A6 = 3 years, A7 = 2 years, A9 = 2 years, A8 = 3 years, A10 = 1 year.

As demonstrated in Table 4, for the confirmatory (core-plus) contrast, pre-intervention assessments for the analytic sample are obtained for most cohorts (B7, A7-A9) in the 2011-12 school year. For most, this was the year prior to the intervention. Cohort B7 is an exception because the schoolwide intervention was active in the 2012-13 school year. Cohort A10 pretest data were obtained for the 2010-11 school year because the AIMS was not administered in the 9th grade. Because there was no schoolwide intervention in 2012-13 in grade 4 and 5, we obtained pretest data for cohorts B6 and C6. Posttest data for all cohorts were gathered in the final year in which testing data are available in the three-year intervention period.

Table 5 outlines the pre and post-intervention assessment data requirements for the exploratory (schoolwide) contrast. Again, for most cohorts (B7, A7-A9) this was obtained from testing data in 2011-12. Cohort A10 uses 8th grade data from the year prior, because 9th grade data are not available. Pre-intervention measurements for cohorts C6 and B6 are obtained from 2012-13 testing data. Posttest data for all cohorts were gathered in the final year in which testing data are available in the three-year intervention period.

A student is selected into the confirmatory (core-plus) study if they were enrolled in the AVID-Core-Plus treatment for one year and they have sufficient pre- and post-intervention data. Once a student is identified as participating in the AVID-Core-Plus treatment, they do not, for the purposes of this study, leave that sample**.** The AVID core program is intended to help students self-motivate, provide them with the tools and skills to increase likelihood of academic success so that these (historically underrepresented) students may go to college. We expect, therefore, that once enrolled in the core program, students should be expected to have these tools to apply in future years. An individual who stops taking the core curriculum may have lower levels of motivation and engagement (which are central foci of the program) and might not perform as well academically as other, more engaged students (and, therefore, might have lower impact estimates). Nevertheless, we have reasoned that this sort of individual-level variation must be addressed in any demonstration of program effectiveness.

<u>College Readiness Outcomes</u>

For college readiness outcomes, we examine two cohorts in a post-only comparison group design. We use only one post-intervention data point in our cross sectional analysis of outcomes. The data were obtained from the CRQ that was administered in the spring of 2015. Only students in cohorts A9 and A10 were included in the samples of both contrasts (see tables above). The students in these cohorts have a maximum possible exposure of three years. This is the same for both the core-plus and schoolwide interventions. Students are assigned to the confirmatory contrast if, on the basis of self-reported data (obtained from the CRQ), they were enrolled for at least one year in the AVID-Core-Plus intervention and they have sufficient outcome data. Students are assigned to the exploratory contrast if self-reported data indicate that they were not enrolled in the AVID-Core-Plus intervention for at least one year and they have sufficient outcome data.

Data Collection for the Evaluation of Impacts

Academic Outcomes

Academic outcomes are measured by rank-based z-score transformations of reading scores (reading domain) and mathematics scores (mathematics domain) on the AIMS and AZMERIT standardized tests that are required for students in grades 3 through 8 and grade 10. We transform scaled scores that we received from the Arizona Department of Education to rank-based z-scores according to the procedures outlined in detail in the *Model Specifications* section. We have elected to transform scores from their scaled score for two reasons: 1) the high school AIMS test (which is taken by students in years prior to 2015) is not on the same scale as the G3-8 test; and 2) the Arizona Department of Education (ADE) transitioned from the AIMS test to a new standardized test, named the AZMERIT, during the final year of the intervention study. As outlined by May (2009), rescaling student-level data to a common metric is an appropriate analytic method to deal with these inconsistencies.[52]

Data collection procedures were consistent in the treatment and comparison conditions. All data collection for academic outcome data is administered by the state and under the responsibility of each school district and the ADE.[53]

College Readiness Outcomes

College readiness outcomes are measured by way of the College Readiness Index (CRI), which is a summative index score of participant responses to a series of items (detailed in Appendix D) included in the *College Readiness Questionnaire* (CRQ). The CRQ was administered to all students in grades 11 and 12 in the treatment and a single comparison district during the spring of 2015. PRG analysts administered the CRQ to all 11th and 12th grade students in attendance at BHS on May 5, 2015. BHS staff proctored questionnaires for students who were absent during the initial administration. PRG staff administered the CRQ to all 11th and 12th grade students in the comparison district on May 6-7, 2015; school staff proctored questionnaires for students who were absent during the initial two-day administration. Students completed paper questionnaires, which were returned to PRG's office for data entry and analysis. Procedures for administration were identical at the treatment and comparison schools.

The CRI is a 17-item index that quantifies the respondent's self-reported college readiness. The index includes items that measure three latent variables (academic preparedness, college knowledge, and college admissions preparedness) that have been identified as predictive of being prepared for college (see College Readiness Item Rationale in Appendix C). Each item/concept included in the index is scored according to the scale scoring protocol provided in Appendix D. The analytic variable represents the total index score for each student. We used Cronbach's alpha to determine the internal consistency (reliability) of all the items included in the scale based on a pre-administration of the instrument in April 2013.

---

[52] May, H., Perez-Johnson, I., Haimson, J., Sattar, S. and 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.

[53] The most recent administration guidelines can be found on the AZED website: http://www.azed.gov/assessment/files/2014/11/azmerit-end-of-course-eoc-guidance.pdf.

The index quantifies the extent to which a student self-reports that he or she is prepared for enrolling in college. The CRI operationalizes the domain of college readiness – that is the degree to which an individual is prepared and able to attend college. Items in the index measure academic preparedness (i.e., content knowledge, academic skills, and academic performance), college knowledge (i.e., understanding of the college admissions process), and college admissions preparedness (i.e., students have taken steps necessary to enroll in college). Face validity is established on the basis that the requirements listed within the index are identified in the literature as a predictor of college attendance and/or as a basic requirement for college eligibility, acceptance, and/or enrollment (see Appendix C for justification of item inclusion and measure construction).

Since the items included in the index are basic requirements for eligibility, acceptance, and attendance in college, the CRI does not suffer from over-alignment with the intervention. Since the same concepts will be measured identically at the treatment and comparison districts the CRI consistently defines "college readiness." The questionnaires are the same for treatment and comparison students, except for the specific names of courses listed in the credit accumulation section, and a series of questions used to identify whether or not the student has taken part in the AVID-Core-Plus intervention. Procedures and protocols are identical. As such the CRQ was consistently collected.

A description of relevant independent variables, along with coding details for covariates, is included in the model specification section.

Model Specifications

Academic Outcomes

We estimate the effects of the intervention on reading and mathematics outcome domains on students in grades 6, 7, 8, and 10. We use a regression-estimated difference-in-difference (DID) approach to model the impacts of the program (Angrist & Pischke, 2009) on the relative performance of students. The DID approach is used frequently in evaluation and education studies as an identification strategy (Schlotter et al., 2010; Murnane & Willett, 2011). The DID model adopted for the analysis of academic outcomes employs a single pre- and post-intervention measurement. A single observation should provide a relatively stable measure of student academic achievement at that point, without the need to average scores across years.

Model Specifications

Here we present the model specification for estimating intervention effects in the confirmatory and exploratory contrasts. The models have been similarly constructed for each outcome/contrast. The only variations are the samples included and contrasts being examined. Separate models are conducted for confirmatory and exploratory contrasts of both mathematics and reading outcomes. The empirical model is estimated with a multi-level model (using Stata command *mixed*). We present the empirical model here, explicitly separating the systems of equations by level:

Level 1:

$$Z_{ijk} = \beta_{0jk} + \beta_{1jk}(TrtYr_{ijk}) + \varepsilon_{ijk}$$

Level 2:

$$\beta_{0jk} = \gamma_{00} + \gamma_{01}(T_{jk}) + \gamma_{0p}(X_{jk}) + \mu_{0jk}$$
$$\beta_{1jk} = \gamma_{10} + \gamma_{11}(T_{jk})$$

Level 3:

$$\gamma_{00} = \pi_{000} + \pi_{001}(SCH_k) + r_{00k}$$
$$\gamma_{01} = \pi_{001}$$

$$\gamma_{0p} = \pi_{00p}$$
$$\gamma_{10} = \pi_{010}$$
$$\gamma_{11} = \pi_{011}$$

Where:

Z_{ijk} – The rank-based z-score transformed scale score in reading (domain 1) and mathematics (domain 2) in the standardized state test (AIMS/AZMERIT) for student *j* in time period *i* in school *k*.

$TrtYr_{ijk}$ – The time-varying indicator of treatment from pre intervention (TrtYr = 0) to post intervention (TrtYr = 1).

T – The treatment indicator for student *j* in school *k*. For the AVID-Core-Plus contrast, T equals 1 if individual has been enrolled in the AVID core program during the intervention period and 0 otherwise. For the schoolwide contrast, T equals 1 if individual has been enrolled in the schoolwide program during the intervention period and 0 otherwise.

X – A vector of time-invariant baseline student-level covariate/controls for student *j* in school *k*. These covariates will include:

(a) Gender (0 = male; 1 = female); variable will be recentered at the grand mean for analysis.
(b) Free or reduced lunch receipt indicator (0 = no; 1= yes); variable will be recentered at the grand mean for analysis.
(c) *Not* Native American indicator (0 = Native American; 1 = not); variable will not be recentered at grand mean for analysis. Nearly all students in the BMS and BHS are Native American. Interpretation of effect estimates will be facilitated by retaining the original coding.
(d) English language learner status (0 = no; 1 = yes); variable will be recentered at the grand mean for analysis.
(e) Special education student status (0 = no; 1 = yes); variable will be recentered at the grand mean for analysis.
(f) Study cohort indicator will be a set of 6-1 = 5 dummy variables to capture the 6 cohorts that moved through the intervention during the study period. Each dummy is initially coded 1 if the individual is in the given cohort and coded 0 otherwise. Dummy variables are then grand mean centered so that the intercept reflects the un-weighted mean cohort effect.

SCH – School-level mean baseline test score in outcome measure (prior to intervention) for students in school *k*. The variable is grand-mean centered so that a value of zero reflects the average mean score for all schools in the analytic sample (see Enders & Tofighi, 2007).[54] Because this variable reflects the average score for the school in which student *i* completes the pretest prior to the treatment/comparison period, the set of *k* schools is not necessarily restricted to the set of schools that are in the treatment or pre-screened comparison districts. Since a student may transfer into (or out of) the treatment and comparison districts during the treatment/comparison periods, they may have completed their pretest in another school.

π_{000} – The estimated pre-treatment mean for the comparison students.

π_{001} – The adjusted mean difference between treatment and comparison students prior to receiving treatment. The pre-intervention mean score for treatment students is estimated by $\pi_{000} + \pi_{001}$.

[54] Enders, C.K., and Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2), 121-138.

π_{010} – The estimated difference in average test scores from pre- to post-intervention for comparison students.

π_{011} – The differences in differences estimate between equivalent group of matched comparison students treatment students. The total estimated mean difference in test scores from pre- to post-intervention for treatment students would be $\pi_{010} + \pi_{011}$.

ε_{ijk} – The random effect at the time-variant individual level (i.e., variation that exists at the pre- *and* post-test). This is the residual variation that remains after all the fixed (or structural) effects in the model and the second (μ_{0jk}) and third-level (r_{00k}) random (or stochastic) effects have been estimated. It is the difference between the test score for individual *j* at time *i* (e.g., pre-test) and the predicted score at each testing occasion.

μ_{0jk} – The variability that remains unexplained by level-2 predictors for individual *j*, across all testing occasions. That is, μ_{0jk} is the deviation of individual *j*'s intercept from the estimated mean baseline status(γ_{00}).

r_{00k} – School-level variability that remains unexplained by level-3 predictor. The deviation of school *k* from the estimated mean baseline status (π_{000}).

Z-Score Transformations

To ensure achievement test data are comparable across grades, we convert individual-level test outcomes (scaled AIMS or AZMERIT test scores) to rank-based z-scores. We perform this conversion separately and state-wide for each test-taking student in each grade level. Z-score transformations are common in educational research (e.g., Gill et al., 2005). We follow the basic procedures outlined in this research, and we detail our approach here.

Since the state transitioned its standardized testing from AIMS to AZMERIT in the 2015 testing year, and the high school tests are on a different scale than the grades 3 through 8 tests, we have decided to transform the scaled scores received from the Arizona Department of Education to rank-based z-scores. The rank-based z-score transposes a student test score into a standardized scaling of student ranks across the state (detailed below).[55] The advantage to this approach is that it estimates achievement effects with fewer assumptions and is, therefore, more robust. As Lockwood writes, "Unless the true scores on the new test are a linear transformation of the true scores on the old test," differences in z-scores may present spurious gains (appendix D, in Gill et al., 2005, p. 201). The rank-based z-score is robust to all (monotone increasing) nonlinear relationships between the new and old test.

For each year of data (both pre-treatment and post-treatment years) our procedure is as follows:

(a) Calculate mean scaled student test score by grade for all students in that grade who have completed the AIMS or AZMERIT for that testing year.
(b) Calculate the standard deviation of scaled student test score by grade for all students in that grade who have completed the AIMS or AZMERIT for that testing year.

---

[55] The rank-based z-score is also referred to as equipercentile equating and is endorsed as an appropriate and robust strategy by May et al., (2005). See also Kolen, M.J., and Brennan, R.L. (2004) *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed). New York: Springer. Gill, B., Hamilton, L.S., Lockwood, J. R., Marsh, J.A., Zimmer, R, Deanna Hill, and Pribesh, S. (2005). Inspiration, Perspiration, and Time: Operations and Achievement in Edison Schools, Santa Monica, Calif.: RAND Corporation, MG-351-EDU, 2005. Available Online: http://www.rand.org/pubs/monographs/MG351.html. May, H., and Supovitz, J.A.. (2006). "Capturing the Cumulative Effects of School Reform: An 11-Year Study of the Impacts of America's Choice on Student Achievement." *Educational Evaluation and Policy Analysis*, 28 (3), 231-257.

(c) Calculate the rank-based z-score based on observed individual scores at each grade level according to the following formula, as outlined by Kirby et al. (2002)[56]:

$$\text{Rank-Based Z score}_i = \text{Probit}^{-1}(r_{ig} / n + 1)$$

where *r* is the rank of student *i* in grade *g* and *n* is the number of students in that grade in the full sample.

College Readiness Outcomes

For both the confirmatory and exploratory contrast, we estimate the effects of the intervention on college readiness outcomes on students in the 11th and 12th grade in the analytic sample. Baseline equivalence is established separately for each using a retrospective measure of parental education, which assesses parent's pre-intervention educational status as the pre-intervention measure (see Appendix C).[57] The design approach for this outcome is a post-only comparison group design, in which we estimate mean post-intervention outcomes for the matched treatment and comparison groups to produce a treatment effect. We estimate this with a multi-level model that takes into effect the nesting of students within schools and accounts for mean school attainment prior to the intervention with a proxy achievement variable.

Model Specification

Now we present the model specification for estimating intervention effects in the confirmatory and exploratory contrasts for the college readiness outcomes. The models are similarly constructed. The only variations are the samples and contrasts being examined. The empirical model is estimated with a multi-level model (using Stata). We present the empirical model here, explicitly separating the systems of equations by level:

Level 1: Students

$$Y_{jk} = \beta_{0k} + \beta_{1k}(T_{jk}) + \beta_{2k}(Grd_{jk}) + \beta_{3k}(X_{jk}) + \varepsilon_{jk}$$

Level 2: School[58]

$$\beta_{0k} = \gamma_{00} + \mu_{0k}$$
$$\beta_{1k} = \gamma_{10}$$
$$\beta_{2k} = \gamma_{20}$$
$$\beta_{3k} = \gamma_{30}$$

Where:

Y_{jk} = College Readiness Scale Score for student *j* in school *k*.

Grd_{jk} = Student grade at time *i* for student *j* in school *k*. The variable will initially be coded as a dummy variable (0 = 11th grade; 1 = 12th grade). We will then re-center the variable at the grand mean.

56 Kirby S.N., McCaffrey D.F., Lockwood J.R., Sloan McCombs J., Naftel S. and Barney H. (2002). "Using state school accountability data to evaluate federal programs: A long uphill road." *The Peabody Journal of Education*, 99(4):122-145.

57 Parental educational status is measured through a series of three questions on the *College Readiness Questionnaire*. Students are asked if either of their parents or guardians finished high school. Then, for the parent who went the furthest in school, students are asked to report a) the level of education completed by the parent, and b) whether or not this has changed over the past three years (which is when the intervention began). For the purpose of establishing baseline equivalence with a pre-intervention measure, responses for parents' current educational attainments are used to compute educational attainment unless students specify that educational status has changed since the beginning of the evaluation period; in this case, retrospective measures are used. Although we do not expect to see much change in parental educational status over the course of the evaluation period, we feel it prudent to consider this as a possibility.

58 Our original model included a variable that captured school-level mean reading test scores. However, because the analysis includes two schools rather than three, there is perfect collinearity between this school-level variable and the treatment indicator. Consequently, we do not include this in the final analytical model.

T_{jk} = The treatment indicator for student *j* in school *k*. For the AVID-Core-Plus contrast, T equals 1 if individual was enrolled in the AVID core program during the intervention period and 0 otherwise. For the schoolwide contrast, T equals 1 if individual was enrolled in the schoolwide program during the intervention period and 0 otherwise.

X = A vector of baseline student-level covariate/controls for student *j* in school *k*. These covariates will include:

(a) Gender (0 = male; 1 = female); variable will be recentered at the grand mean for analysis.
(b) Household Resources (continuous; values range from 0= no resources reported to 4= all resources reported); variable will be recentered at the grand mean for analysis.
(c) *Not* Native American indicator (0=Native American; 1 = not);
(d) Age (continuous); variable will be recentered at the grand mean for analysis.
(e) Living with both biological parents (0=no; 1= yes); variable is recentered at the grand mean for analysis.
(f) Number of siblings (continuous); variable will be recentered at the grand mean for analysis.
(g) Parents' education (continuous; values range from 1 to 6); this is the variable used to establish baseline equivalence. Prior research establishes parents' educational attainment as one of the most salient predictors of whether or not students enroll and succeed in college (see the College Readiness Questionnaire Item Rationale in Appendix C). Variable is recentered at the grand mean for analysis.

γ_{00} = The estimated mean CRI score for the comparison students, controlling for other variables in the model.

γ_{10} = The estimated mean CRI score "impact" for the treatment students, controlling for other variables in the model. The estimated mean college readiness score for treatment students can be obtained by adding the two parameter estimates $\gamma_{00} + \gamma_{10}$.

ε_{jk} – The residual variation that remains at the individual level after all the fixed (or structural) effects in the model and the random (or stochastic) effects at the second level (μ_{0k}) have been estimated.

μ_{0k} – The un-modeled school-level variability that is unexplained by level-2 predictors. That is, μ_{0k} is the deviation of the school's mean CRI score from the estimated mean CRI score for comparison students (γ_{00})and the variation is explained by school-level predictors (γ_{01}).

Decision Rules for Inclusion/Exclusion of Covariates

All covariates identified in the *Model Specifications* section will remain in the empirical models as indicated, regardless of statistical significance, because they are theoretically important control variables.

Treatment of Missing Data

We do not impute missing outcome data, as per WWC evidence standards. Also we do not impute missing covariate data. Impact analyses samples include only those observations that have non-missing posttest, and in the case of academic outcomes pretest, data

Calculation of Effect Size

We calculate effect sizes in accordance with the guidelines published in the *WWC Procedures and Standards Handbook*, Version 3.0. The standard deviation for each outcome is an estimated value from the data. We use a pooled standard deviation. We outline the formulas for each outcome below.

Academic Outcomes

Standardized effect sizes for each contrast (exploratory and confirmatory) within each domain (math and reading) are estimated with the standardized mean difference, known as *Hedges' g*:

$$g = \frac{\pi_{011}}{\sqrt{\frac{(n_t - 1)S_t^2 + (n_c - 1)S_c^2}{(n_t + n_c - 2)}}}$$

where: π_{011} is the "differences in differences" estimate of the difference between comparison and treatment students that is produced by the equation detailed in the *Model Specifications* section, and n_t and n_c are the sample sizes, and S_t and S_c are the student-level standard deviations for the analytic treatment and comparison groups, respectively. We produce separate effect size calculations for each domain (reading and mathematics achievement) within each contrast (confirmatory and exploratory).

College Readiness Outcomes

Standardized effect sizes for each contrast (exploratory and confirmatory) are estimated with the standardized mean difference, known as *Hedges' g*:

$$g = \frac{\gamma_{10}}{\sqrt{\frac{(n_t - 1)S_t^2 + (n_c - 1)S_c^2}{(n_t + n_c - 2)}}}$$

where: $\gamma_{10}$is the estimated difference in mean CRI scores between comparison and treatment students that is produced by the equation detailed in the *Model Specifications* section, and n_t and n_c are the sample sizes, and S_t and S_c are the student-level standard deviations for the analytic treatment and comparison groups, respectively. We produce separate effect size calculations for each domain (reading and mathematics achievement) within each contrast (confirmatory and exploratory).

Methods for Baseline Balance Testing

Verification of Baseline Equivalence on Observed Characteristics

Per NEi3 guidance, a QED "must provide evidence that the groups being contrasted (e.g., the treatment and comparison groups) are equivalent at baseline on a pre-intervention measure of the outcome variable or on a variable or variables that are highly correlated with the outcome measure" (p. 14). For academic outcomes, the WWC also provides some specific direction on minimum requirements on which pre-intervention measures should be used to establish baseline equivalence.[59] We follow these guidelines when they are available for our outcomes (i.e., academic outcomes). We include the measures used to establish baseline equivalence in our impact estimation models, regardless of the magnitude of difference, because it adds precision to our estimates. As is indicated elsewhere in this plan and in the contrast tool, we establish baseline equivalence with the following variables:

Academic Outcomes:

1. AIMS reading pre-test (a pre-intervention outcome measure) for both contrasts in reading domain.

[59] The WWC mathematics protocol states that baseline equivalence must be established (at least on) a pretest of the outcome measure and grade level. As is noted in the contrasts section, at least one of these measures is included in as the means to establish baseline equivalence.

2. AIMS mathematics pre-test (a pre-intervention outcome measure) for both contrasts in mathematics domain.

College Readiness Outcome: Parents' education status (an SES indicator and highly-correlated variable) for both contrasts.

Model Specifications for Tests of Baseline Equivalence
To test for baseline equivalence in accordance with the NEi3 guidance, first we produce model-based estimates of the difference between the treatment and comparison groups for the identified baseline equivalence variable for each contrast. We then compute the pooled standard deviation of these variables. Finally, for each contrast, we produce a standardized difference of means by dividing the first term by the second. We outline a three-step procedure for each outcome below.

Academic Outcomes

Step 1. First we present the model specification for estimating the difference between treatment and comparison groups on the pre-intervention measures. The same procedures are followed for both the confirmatory and exploratory contrast – the only difference being the samples included and contrasts being examined. Separate models are run for mathematics and reading achievement outcomes, but the models themselves are identical, except, as noted below, for the dependent variable. The DID structure allows us to estimate the adjusted mean difference in the pre-intervention measure, with a model that is similar to the one we use to estimate outcomes, though we exclude demographic covariates in this case. The empirical model is estimated with a multi-level model (using Stata). We present the empirical model here, explicitly separating the systems of equations by level:

Level 1:

$$Z_{ijk} = \beta_{0jk} + \beta_{1jk}(TrtYr_{ijk}) + \varepsilon_{ijk}$$

Level 2:

$$\beta_{0jk} = \gamma_{00} + \gamma_{01}(T_{jk}) + \mu_{0jk}$$
$$\beta_{1jk} = \gamma_{10} + \gamma_{11}(T_{jk})$$

Level 3:

$$\gamma_{00} = \pi_{000} + r_{00k}$$
$$\gamma_{01} = \pi_{001}$$
$$\gamma_{10} = \pi_{010}$$
$$\gamma_{11} = \pi_{011}$$

Where:

Z_{ijk} – The z-score transformed scale score in reading (domain 1) and mathematics (domain 2) in the standardized state test (AIMS/AZMERIT) for student *j* in time period *i* in school *k*. Time period *i* will either be the pre- or posttest.

$TrtYr_{ijk}$ – The time variant indicator of treatment from pre-intervention (TrtYr = 0) to post-intervention (TrtYr = 1).

T – The group treatment indicator for student *j* in school *k*. For the AVID-Core-Plus contrast, T equals 1 if individual was enrolled in the AVID core program during the intervention period and 0 otherwise. For the schoolwide contrast, T equals 1 if individual was enrolled in the schoolwide program during the intervention period and 0 otherwise.

π_{000} – The estimated pre-treatment mean for the comparison students.

π_{001} – The adjusted mean difference in the baseline equivalence (pretest) variable between treatment and comparison students prior to receiving treatment. This is the estimate that becomes the numerator in the final equation to produce the standardized mean difference in the reading pre-test (domain 1) or math pre-test (domain 2).

We exclude descriptions of the other parameter estimates that are not relevant to establishing baseline equivalency. Their interpretation, however, remains identical to what we present in the *Model Specifications* section. We produce these estimated differences on the identified pre-intervention measure for each domain (reading and mathematics achievement) within each contrast (confirmatory and exploratory).

Step 2. Compute the pooled standard deviation of the pre-intervention measures used to establish baseline equivalence as follows:

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

Where: n_t and n_c are the sample sizes, and S_t and S_c are the student-level standard deviations for the pre-intervention measures for the analytic treatment and comparison groups, respectively. We produce separate calculations of the pooled standardized deviation for each domain (reading and mathematics achievement) within each contrast (confirmatory and exploratory).

Step 3. Produce the standardized difference of means as follows:

$$g = \frac{\pi_{001}}{S_p}$$

Where: π_{001} is the adjusted mean difference in the variable selected to establish baseline equivalence (pre-test in reading and mathematics) for the treatment and comparison groups (calculated in Step 1) and S_p is the pooled standard deviation (produced in Step 2).[60]

The standardized mean difference is used to establish baseline equivalence of the treatment and comparison samples in accordance with the guidelines in the NEi3 *Analysis and Reporting Plan*, Version 1.0 (p.15). According to these guidelines, assessments of the magnitude of difference between groups are based on the following rules:

- less than or equal to .05 standard deviations – equivalence has been established;
- greater than .05 standard deviations but less than or equal to .25 standard deviations and variable included in the model – equivalence has been established;[61]
- greater than .25 standard deviations – baseline equivalence not established.

[60] This comparison excludes cases with missing values, as per NEi3 review guidelines (NEi3 *Analysis and Reporting Plan*, Version 1.0, p.14).
[61] As noted in the model specifications section, we intend to include the variable used to establish baseline equivalence in the regression model, regardless of the magnitude of difference, to improve the precision of our estimates.

College Readiness Outcomes

Step 1. As with academic and dropout outcomes, we first present the model specification for estimating the difference between treatment and comparison groups on the identified pre-intervention measure. The same procedures are followed for both the confirmatory and exploratory contrast – the only difference being the samples included and contrasts being examined. The empirical model is estimated with a multi-level model (using Stata). We present the empirical model here, explicitly separating the systems of equations by level:

Level 1: Students

$$Y_{jk} = \beta_{0k} + \beta_{1k}(T_{jk}) + \beta_{2k}(Grd_{jk}) + + \varepsilon_{jk}$$

Level 2: School

$$\beta_{0k} = \gamma_{00} + \mu_{0k}$$
$$\beta_{1k} = \gamma_{10}$$
$$\beta_{2k} = \gamma_{20}$$
$$\beta_{3k} = \gamma_{30}$$

Where:
Y_{jk} = Parents' education status (continuous; values range from 1 to 6). Prior research establishes parents' educational attainment as one of the most salient predictors of whether or not students enroll and succeed in college (see the College Readiness Questionnaire Item Rationale in Appendix C).

Grd_{jk} = Student grade at time *i* for student *j* in school *k*. The variable is initially coded as a dummy variable (0 = 11th grade; 1 = 12th grade). We then re-center the variable at the grand mean.

T_{jk} – The treatment indicator for student *j* in school *k*. For the AVID-Core-Plus contrast, T equals 1 if individual was enrolled in the AVID core program during the intervention period and 0 otherwise. For the schoolwide contrast, T equals 1 if individual was enrolled in the schoolwide program during the intervention period and 0 otherwise.

γ_{10} – The estimated mean difference in parents' education between treatment and comparison students. This is the parameter estimate of interest for baseline equivalence. This estimate represents the adjusted mean difference in the baseline equivalence variable (Y_{jk}) between treatment and comparison students prior to receiving treatment.

We exclude descriptions of the other parameter estimates that are not relevant to establishing baseline equivalence. We produce these estimated differences on the identified pre-intervention measure used to establish baseline equivalence for each contrast (confirmatory and exploratory).

Step 2. Compute the pooled standard deviation of the pre-intervention measures used to establish baseline equivalence as follows:

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

Where: n_t and n_c are the sample sizes, and S_t and S_c are the student-level standard deviations for the pre-intervention measures for the analytic treatment and comparison groups, respectively. We produce separate calculations of the pooled standardized deviation for both the confirmatory and exploratory contrasts.

Step 3. Produce the standardized difference of means as follows:

$$g = \frac{\gamma_{10}}{S_p}$$

Where: γ_{10} is the estimated mean difference in the identified variable used to establish baseline equivalence (parents' education status) between the treatment and comparison groups (calculated in Step 1) and S_p is the pooled standard deviation (produced in Step 2).[62]

The standardized mean difference is used to establish baseline equivalence of the treatment and comparison samples in accordance with the guidelines in the NEi3 *Analysis and Reporting Plan*, Version 1.0 (p.15). According to these guidelines, assessments of the magnitude of difference between groups are based on the following rules:

- less than or equal to .05 standard deviations – equivalence has been established;
- greater than .05 standard deviations but less than or equal to .25 standard deviations and variable included in the model – equivalence has been established;[63]
- greater than .25 standard deviations – baseline equivalence not established.

We calculate the standardized mean difference between the treatment and comparison sample for each domain (reading and mathematics achievement) within each contrast (confirmatory and exploratory).

[62] This comparison excludes cases with missing values, as per NEi3 review guidelines (NEi3 Guide, p.14).
[63] As noted in the model specifications section, we include the variable used to establish baseline equivalence in the regression model, regardless of the magnitude of difference, to improve the precision of our estimates.

Appendix B: Z-Score Transformations

To ensure achievement test data are comparable across grades, we convert individual-level test outcomes (scaled AIMS or AZMERIT test scores) to rank-based z-scores. We detail these procedures in the *Model Specifications* section of Appendix A. Below, we provide an overview of the procedures, and we present tables of z-scores and corresponding standardized test scores for math and reading tests during the study period (Tables B1 and B2). These transformations allow us to compare achievement across years and changes in standardized tests.

Since the state transitioned its standardized testing from AIMS to AZMERIT in the 2015 testing year, and the high school tests are on a different scale than the grades 3 through 8 tests, we transform the scaled scores received from the Arizona Department of Education to rank-based z-scores. We perform this conversion separately and state-wide for each test-taking student in each grade level. The rank-based z-score transposes a student test score into a standardized scaling of student ranks across the state (detailed below).

For each year of data (both pre-treatment and post-treatment years) our procedure is as follows:

(a) Calculate mean scaled student test score by grade for all students in that grade who have completed the AIMS or AZMERIT for that testing year.
(b) Calculate the standard deviation of scaled student test score by grade for all students in that grade who have completed the AIMS or AZMERIT for that testing year.
(c) Calculate the rank-based z-score based on observed individual scores at each grade level according to the formula outlined by Kirby et al. (2002).[64]

[64] Kirby S.N., McCaffrey D.F., Lockwood J.R., Sloan McCombs J., Naftel S. and Barney H. (2002). Using state school accountability data to evaluate federal programs: A long uphill road. *The Peabody Journal of Education*, 99(4):122-145.

Table B1. Raw Standardized Math Test Scores and Z-Scores by Grade and Year

		Raw AIMS (2011-2014) and AZMERIT (2015) Math Scores				
Math[65]	**Z-score**	**2011**	**2012**	**2013**	**2014**	**2015**
Grade 4						
	1	437	437	434	433	3590
	0	386	388	385	382	3553
	-1	334	338	336	331	3515
Grade 5						
	1	447	447	445	444	3624
	0	397	398	398	397	3586
	-1	348	350	350	349	3547
Grade 6						
	1	465	464	461	460	3651
	0	413	414	414	413	3615
	-1	360	365	368	366	3580
Grade 7						
	1	476	484	482	477	3669
	0	426	431	432	430	3632
	-1	375	378	382	383	3595
Grade 8						
	1	483	489	489	488	3698
	0	435	439	440	440	3660
	-1	386	388	390	392	3622
Grade 10						
	1	548	547	544	545	n/a
	0	501	501	501	503	n/a
	-1	454	456	458	461	n/a

[65] 2011-2014 scores are from the AIMS test; 2015 scores are from the AZMERIT test.

Table B2. Raw Standardized Reading Test Scores and Z-Scores by Grade and Year

		Raw AIMS (2011-2014) and AZMERIT (2015) Reading Scores				
Reading[66]	**Z-score**	**2011**	**2012**	**2013**	**2014**	**2015**
Grade 4						
	1	534	528	528	524	2543
	0	484	482	484	482	2514
	-1	435	437	440	441	2485
Grade 5						
	1	547	548	549	545	2556
	0	502	503	505	504	2528
	-1	457	459	461	463	2500
Grade 6						
	1	558	561	559	557	2571
	0	515	516	515	515	2541
	-1	472	471	472	472	2511
Grade 7						
	1	577	582	585	577	2577
	0	531	536	538	536	2548
	-1	486	490	492	494	2519
Grade 8						
	1	581	577	577	573	2588
	0	527	526	526	525	2559
	-1	473	476	476	476	2530
Grade 10						
	1	762	761	756	766	2596
	0	712	715	715	721	2569
	-1	661	669	675	676	2543

[66] 2011-2014 scores are from the AIMS test; 2015 scores are from the AZMERIT test.

Appendix C: *College Readiness Questionnaire* Item Rationale

College Readiness

College readiness is often conceptualized according to Conley's (2007:3) definition, "as the level of preparation a student needs in order to enroll and succeed – without remediation – in a credit-bearing general education course at a post-secondary institution that offers a baccalaureate degree or transfer to a baccalaureate program." Though, in practice, there is variation in what researchers emphasize as the core components of college readiness and even more variation in how they operationalize these components. For our study, we have created a college readiness index that includes two domains that are consistent across the literature, plus a third that we added due to its relevance for student outcomes. The present conceptualization draws on the work of Gurantz and Borsato (2012) and encapsulates most of the domains mentioned in other research on college readiness.

For Gurantz and Borsato (2012:7), two of the primary domains or indicators of college readiness are academic preparedness and college knowledge.[67] Gurantz and Borsato prefer these indicators because they have "a consistent and predictable relationship with CR [College Readiness]," and they are actionable, meaning they can be affected by the educational system. In addition, we also include a measure of college admissions preparation which we discuss the importance of below.

Academic Preparedness

Academic preparedness refers to "key academic content knowledge and cognitive strategies needed to succeed in doing college-level work" (Gurantz and Borsato 2012:7). In practice, this domain is often subdivided. For instance, Conley (2007) breaks it down into *key content knowledge* and *key cognitive strategies,* and Roderick et al. (2008; 2009) discuss this as *content knowledge* and *core academic skills*; further, Gheen et al. (2012) refer to this as *academic and technical performance,* and *engagement.* As Roderick et al. (2009) acknowledge, the difference in these concepts is subtle; therefore, our decision to recognize them under one domain will facilitate operationalization and interpretation. There are a number of ways to operationalize the concept of academic preparedness. Since this is a multifaceted concept, we employ a number of measures of preparedness: minimal course requirements, enrollment in rigorous courses, and enrollment in Algebra II or higher math course (items listed below).[68] All of these indicators have individually, and as elements in indices, been shown to be strongly correlated with academic achievement and educational attainment.

Minimal course requirements refers to whether or not students are on track to secure high school graduation and college qualifications associated with course requirements; this is included as it is a measure of whether or not students have enrolled in the minimum course work necessary to graduate from high school and be accepted into college (Cabrera and La Nasa 2001; Roderick et al 2009).[69] Rigorous

[67] In their work on college readiness, Gurantz and Borsato identify and measure a third domain of college readiness – academic tenacity – which we are unable to capture. According to Gurantz and Borsato (2012), academic tenacity refers to beliefs and attitudes that underlie achievement and educational attainment. This is similar to what other researchers refer to as *non-cognitive skills* (Conley 2007; Roderick et al 2009) or *habits of the mind* and *academic behaviors* (Conley 2007). Due to data collection constraints, we are unable to follow Gurantz and Borsato's (2012) suggestion and use attendance and disciplinary infractions as proxies for tenacity. Therefore, we focus on academic preparedness and college knowledge.

[68] GPA is often used as a measure of academic preparedness for a number of reasons, including that it is used by colleges in the admissions process; it is indicative of whether or not students have mastered course material, and most importantly, it is a strong predictor of enrollment in a four-year college and completion of college (within six years) (Adeleman 2006; Cabrera and La Nasa 2001; Farrington et al 2012; Gurantz and Borsato 2012; Roderick et al 2009). Though GPA is considered an important indicator of preparedness, we do not include it in our *College Readiness Index* because GPA may not be "collected in the same manner" (What Works Clearinghouse 2013: 15) at the treatment and comparison schools. That is, grades may not be assigned in the same manner across classrooms or schools, and GPA may not be calculated in the same way. So comparing GPAs may produce unreliable estimates of programmatic impact.

[69] Our measures of minimal course requirements are informed by the Arizona Department of Education graduation requirements and by the Arizona Board of Regents Tri-University Admission requirements. We only include in the index course requirements in

coursework differs from course completion in that it refers to students enrolling in academically challenging classes – for our purposes Advanced Placement (AP) and Dual Enrollment courses. Enrollment in such classes is considered by some to be the most salient predictor of college enrollment and completion (Adeleman 2006; Cabrera and La Nasa 2001; Cha and Reinert 2010; Gheen et al 2012; Gurantz and Borsato 2012; Watt et al 2007). As an additional measure of rigorous course work, we also include a measure of whether or not a student has enrolled in Algebra II (or a higher math course). This has been shown to be especially correlated with high school graduation, college enrollment, and college completion (Adelman 2006; Cha and Reinert 2010; Long, Iatarola, & Conger, 2009; Watt et al 2007).

College Knowledge

College knowledge refers to the contextual knowledge of college admissions and skills necessary in order for a student to successfully navigate the admissions process (Conley 2007; Gurantz and Borsato 2012; Roderick et al 2009). When conceptualizing college or contextual knowledge, researchers often highlight the importance of students understanding the admissions process in terms of application timelines, testing requirements, and financial aid options and also understanding the different college options (Chait et al 2009; Conley 2007; Guarantz and Borsato 2012; Lundell et al 2004; Watt et al 2007). Though Conley is developing a "College Ready School Diagnostic" which will include measures of contextual skills and awareness (Smaller Learning Communities Program 2011), a systematic tool for measuring student college knowledge is not yet available.

College Admissions Preparation

In addition to academic preparedness and college knowledge, we also capture a third domain in our college readiness index – that of college admissions preparation. The items included under this domain are those that show that students have a contextual understanding of the admissions process and have taken the steps necessary to enroll in college, measures include: completion of a FAFSA; completion of other financial aid and scholarship applications; completion of college applications; visits to college campuses; discussions about the admissions process; and completion of college entrance exams. Therefore, in order to assess these three key components of college readiness, we developed the *College Knowledge Index* (items listed below).

College Readiness Items

Academic Preparedness (developed by PRG and adapted from High School Longitudinal Survey and National Educational Longitudinal Survey)

14. Have you taken or are you currently enrolled in: English Courses (listed)
15. Have you taken or are you currently enrolled in: Mathematics Courses (listed)
16. Have you taken or are you currently enrolled in: Science Courses (listed)
18. Have you taken or are you currently enrolled in any Dual Enrollment courses (either online or in your school?

College Knowledge (developed by PRG)

34. All four-year colleges and universities have the same admission requirements.
35. Almost all four-year colleges and universities require you to take three or more years of mathematics classes that are Algebra I and higher, while in high school.
36. Many junior or community colleges only require that you have a high-school diploma or GED.
37. Many junior or community colleges have late application deadlines.

---

the subjects of English, Mathematics, and Science, as these are the same for both graduation and college admissions and both comparison and treatment schools offer the courses to fulfill these requirements. See http://www.azed.gov/ for information on high school graduation and college admission requirements.

College admissions preparation (developed by PRG)

24. Have you ever taken the ACT?
25. Have you ever taken the SAT?
26. Have you ever completed and submitted the Free Application for Federal Student Aid (FAFSA)?
27. Have you ever applied for a college scholarship or another type of financial aid to pay for college (other than the FAFSA)?
28. Have you ever visited a community college, college, or university campus with the intention of possibly attending at some point in the future?
29. Have you ever completed and submitted a community college, college, or university application?
32. Have you ever had a discussion with a counselor, teacher, or other adult at your school to get information or advice on the college application process, requirements, or going to college?
33. Have you ever talked with your parents to get information or advice on the college application process and requirements, or about going to college?

Background Information

We include in the questionnaire items that assess basic demographic characteristics that describe the respondent and act as covariates in the analysis. These characteristics are also used for matching treatment school cases to comparison school cases. The actual date of birth has not been included. This is due to the request of school district officials.

1. What is your gender?
2. What grade are you currently in?
3. In what year were you born?
4. What is your age?
5. Please select one or more of the following that best describes your race?
6. Are you Hispanic or Latino/Latina?

Family Structure

There is a great deal of research that examines the effect of family structure on individual well-being. For the current evaluation, most important is the body of research that explores the relationship between family structure and academic and schooling outcomes. Largely, this research has found that individuals who live in nuclear families with both biological parents have better academic/schooling outcomes as compared to individuals who live in single-parent or step-parent families (Astone and McLanahan 1991, 1994; Biblarz and Raftery 1999; Dawson 1991; Ginther and Pollak 2004; Hamilton, Cheng, and Powell 2007; Jeynes 2005; Rosenfeld 2010). Some research has also shown that that children of two adoptive parents (Hamilton et al 2007) and same-sex couples (Rosenfeld 2010) also have more positive academic outcomes, in terms of progress through school, as compared to single-parent or step-parent families. Regarding outcomes, research has shown that children in single-parent or step-parent families have lower grade-point averages (Astone and McLanahan 1991) and lower test scores (Jeynes 2005); they are more likely to have to repeat a grade (Astone and McLanahan, 1994; Dawson 1991); they are more likely to drop out of high school (Astone and McLanahan, 1994; Ginther and Pollack 2004); and they have lower educational attainment (fewer years of schooling, less likely to attend or graduate from college) (Biblarz and Raftery 1999; Ginther and Pollack 2004).

In addition to the literature on the parental structure of the family, there is also research that suggests that the number of siblings a child has affects educational outcomes. Number of siblings has been shown to be negatively associated with family resources, such as financial support for college (Roscigno and Ainsworth-Darnell 1999; Steelman & Powell, 1989), years of schooling, high school graduation, and college attendance (Biblarz and Raftery 1999; Ginther and Pollack 2004). That is, the more siblings a child has, the lower his or her educational attainment.

We include two measures of family structure. The first captures parental structure of the family. This measure is taken from the NELS, which was used in Jencks (2005) study. It also provides categories similar to those used by Astone and McLanahan (1994) and Biblarz and Raftery (1999).

Family Structure (Adapted from Educational Longitudinal Survey)

7. At this time, with whom do you live? Mark all that apply.
8. How many brothers and sisters do you have? Please include any stepbrothers and/or stepsisters if they live or have lived in your home.

Socioeconomic Status and Parental Education

Socioeconomic status (SES) is a multidimensional concept that is essentially a measure of economic wellbeing and social status. According to Sirin (2005: 418), “It describes an individual’s or family’s ranking on a hierarchy according to access to or control over some combination of valued commodities such as wealth, power, and social status.” SES generally is conceptualized as comprising three primary elements: educational attainment, occupation, and income of householder(s) (Ainsworth 2002; Boudreau and Poulin 2009; Currie et al 1997; Entwisle, D. & Astone 1994; Sirin 2005). In addition, some scholarship also recognizes a fourth element: home resources, such as books, computers, and tutoring (Sirin 2005). Family socioeconomic status is important because it in large part dictates what resources (human, social, and economic capital) children have at their disposal (Coleman 1988; Sirin 2005). Children from higher SES families (that is, those with more resources) fare better in school (are more likely to turn in homework, get better grades, do better on achievement tests) (Ainsworth 2002; Roderick et al 2009; Sirin 2005); they are more likely to take rigorous courses in high school, such as college prep courses or AP courses (Roderick et al 2009); they are more likely to acquire college qualifications and graduate from high school (Cabrera and La Nasa 2001); they are more likely to plan to go to, apply to, and enroll in college (Cabrera and La Nasa 2001; Perna and Titus 2007; Sewell and Shah 1967); and they achieve higher levels of education (Biblarz and Raftery 1999; Sewell and Shah 1967; Sirin 2005). Several meta-analyses have been conducted which show that family SES is, in terms of effect size, moderately to strongly correlated with academic achievement and educational attainment; moreover, when included in multivariate models, it is also one of the strongest predictors of academic outcomes (Sirin 2005).

Though SES is clearly a predictor of academic achievement and college readiness and should, therefore, be included in our analyses, operationalization is an issue. Measuring income is problematic because it is a sensitive topic, and adolescents often are unaware of their family’s income; therefore, it often leads to a high degree of nonresponse (Boudreau and Poulin 2009; Ensminger et al 2000; Entwisle, D. & Astone, 1994). Children have been shown to have knowledge of mother’s employment status (i.e., working full time, part time, not working), but less of father’s employment status (Ensminger et al 2000); however, occupation is prone to both nonresponse and reporting error (Boudreau and Poulin 2009; Currie et al 1997; Ensminger et al 2000; Entwisle, D. & Astone, 1994). Parental education is considered a better measure of SES in part because it is moderately to strongly correlated with the other measures of SES (Sirin 2005; Ensminger 2000), and it does not have the problems of reliability and nonresponse that are associated with income and occupation (Currie et al 1997; Ensminger et al 2000; Entwisle, D. & Astone).

Considering these issues, we do not ask students about parental income or occupation, but instead focus on parental educational level and household resources.

Education is a particularly important element of SES; researchers have found parents' education is positively associated with educational attainment (i.e., number of years of schooling, high school graduation, and college attendance) (Biblarz and Raftery 1999; Bradley and Corwyn 2002; Ginther and Pollack 2004; Sirin 2005). In a review of studies on the link between parental education and children's educational attainment, Choy (2001) of the U.S. Department of Education, National Center for Education Statistics indicates that parental education is perhaps the most salient predictor of whether or not students enroll, persist in, and graduate from college. According to Choy, students whose parents did not go to college and those whose parents did not graduate from high school "are at a distinct disadvantage when it comes to postsecondary access—a disadvantage that persists even after controlling for other important factors such as educational expectations, academic preparation, support from parents and schools in planning and preparing for college, and family income." In part, this is related to the fact that, in the U.S., education and income are highly correlated, so that children of parents who are more highly educated are more likely to have access to resources afforded by better economic wellbeing, such as better schooling options, educational resources (e.g., computers, books) in the home (Bradley and Corwyn 2002; Orr 2003; Sirin 2005). However, this is also a result of intangible resources provided by parents' education. For instance, parents with higher educations are a knowledge resource for their children – they can help them with school work, help them make college decisions, and help them through the application process (Bradley and Corwyn 2002; Coleman 1988; Rowan and Kenyon 2008). In order to measure parental education, we ask students to indicate the highest level of education that their parent who has gone the furthest in school has attained.

In addition, we use an alternate measure of SES: home resources, which might better capture the diversity of resources and support for schooling in the population. [70] Home resources have been shown to be positively associated with academic outcomes including higher GPAs, higher standardized test scores, and higher educational attainment (Orr 2003; Roscigno and Ainsworth-Darnell 1999; Sirin 2005).

Socioeconomic Status (adapted from High School Longitudinal Survey and National Educational Longitudinal Survey)

9. Which of the following does your family have in your home: a) a computer; b) a dictionary; c) a calculator; d) more than 25 books
10. Did either your mother or father (or one of your guardians if you don't live with either parent) finish high school? It's okay if you are not 100% certain, please give us your best guess.
11. For the next question, we want to know about your parent (or guardian) who went the furthest in school. We have listed a number of educational diplomas or degrees people can earn. Please check each box that you know or are pretty sure this parent (or guardian) received.
12. Thinking again about your parent or guardian who went the furthest in school, has his or her education level CHANGED in the past three years? *(For example, received a GED, or an Associate's or Bachelor's degree in the past three years?)*

[70] It is notable that, according to Census Bureau estimates (2010 ACS 3-yr estimates), Native American/American Indian populations typically do not attain high levels of education – approximately a quarter (24%) of Native American/American Indian individuals do not graduate from high-school, and a smaller proportion earns at least a bachelor's degree (13%). Correspondingly, Native American/American Indian populations also earn low incomes – median income was approximately $20,000 less than the national median, and the percent of families in poverty was over 10% higher (22% vs 11%). Therefore, we expect our study population to be relatively homogenous (and disadvantaged) in terms of financial and educational aspects of SES.

Appendix D: College Readiness Index Scoring Protocol

College Readiness Index – 17 point scale

Academic Preparedness – 5 points possible

14-16: Have you taken or are you currently enrolled in: (English, Math, Science courses listed)
18. Have you taken or are you currently enrolled in any Dual Enrollment courses (either online or in your school?

Minimum Course Requirements – 3 points possible
Courses taken or enrolled in

- 4 English courses = 1 point; l partial points
- 3 Math courses = 1 point; l partial points
- 3 Science courses = 1 point; l partial points

Rigorous Course work – 2 points possible
Courses taken or enrolled in

- Algebra II or higher = 1 point
- Dual Enrollment courses= 1 point

College Knowledge – 4 points possible

34. All four-year colleges and universities have the same admission requirements.
35. Almost all four-year colleges and universities require you to take three or more years of mathematics classes that are Algebra I and higher, while in high school.
36. Many junior or community colleges only require that you have a high-school diploma or GED.
37. Many junior or community colleges have late application deadlines.

College Knowledge Index – 4 points possible

- All questions answered correctly = 4 points; partial points allotted (questions equal weight)

College admissions preparation – 8 points possible

24. I have taken the SAT at least one time
25. I have taken the ACT at least one time

Completion of college entrance exams – 2 points possible

- Taken SAT = 1 point
- Taken ACT = 1 point

26. I have completed and submitted the Free Application for Federal Student Aid (FAFSA)
27. I have applied for at least one college scholarship (or another type of financial other than FAFSA)

Completion of Financial Aid and Scholarship Applications-2 points possible

- Completed FAFSA = 1 point
- Applied for scholarships or other financial aid = 1 point

28. I have visited at least one community college, college, or university campuses with the intention of possibly attending at some point in the future

Visited college campuses – 1 point possible

- Visited at least one campus = 1 point

29. I have completed and submitted at least one community college, college, or university application

Applied to colleges – 1 point possible

- Applied to at least one college = 1 point

32. I have had at least one discussion with a counselor, teacher, or other adult at my school to get information or advice on the college application process, requirements, or going to college.
33. I have had at least one discussion with my parents to get information or advice on the college application process, requirements, or going to college

Discussed admissions process – 2 points possible

- Discussed admissions process with school faculty = 1 point
- Discussed admissions process with parents or guardians = 1 point

Appendix E: Data Management

Arizona Department of Education Data

In this section we detail our data management procedures for math and reading test and background characteristic data that we received from ADE. The datasets ADE provided are divided by testing year with unique observations for math scores and reading scores. Some of the background information for students was different for their math observation than their reading observation; therefore, we determined to create separate datasets for math test results and reading test results. Observations are identified with a masked student ID.

We requested and received datasets from ADE including statewide math and reading scores and demographic data for all students in grades 4-6, 8, and 10 for 2011-2015. We merged data from year to year using masked student ID numbers provided by ADE. We verified that individuals matched from year to year using demographic characteristics that we expect to remain the same (e.g., gender, ethnicity). During data checks, we found numerous instances of inconsistencies in gender and ethnicity for students from year to year. ADE reported that they are confident that the masked IDs that they provided to us link the same student across years and across AIMS and AZMERIT databases. ADE resubmitted the datasets, and the year-to-year inconsistencies were reduced to 2 for gender and 32 for ethnicity. We asked the ADE analyst who provided the data to look into each of the cases. The analyst confirmed that, for the two gender inconsistencies, the name is the same in both cases. The analyst looked into each of 32 cases of inconsistencies in ethnicity and provided a spreadsheet with the ethnicity from the ADE database. We use the data provided by ADE for analysis.

The data included several duplicate student ID numbers where background characteristics and grade level do match and appear to be different students. In these cases, we dropped the observations. In cases where we have duplicate student IDs and their background characteristics match, ADE informed us that it is likely that the student was allowed to take a test in their grade level and another grade level. In these cases, we keep data from the test that is the same level as their current grade.

AZMERIT Math Test

In 2015, ADE moved from the AIMS test to the AZMERIT test. The AZMERIT test no longer has end-of-grade level tests and, instead, has subject matter tests (e.g., Algebra, Geometry). In order to align 2015 data with previous years, we do not include math tests that are not grade-level specific. Therefore, we dropped all observations associated with Algebra I, Algebra II, and Geometry tests. As a result, we do not have posttest data on math achievement for students in 10^{th} grade in 2015.

Missing Data

We do not impute missing outcome data, as per WWC evidence standards. We also do not impute missing covariate data. Case-wise deletion is used for the cases with missing outcome data and the very few instances where there was sufficient outcome data but missing covariate data.

Appendix F: Baseline Equivalence of Samples

Academic Outcomes

Table F1. Background Characteristics of AVID-Core-Plus Students and Comparison Students at Baseline – Mathematics

Characteristic	AVID-Core-Plus Students (*n* = 262)	Comparison Students (*n* = 262)	Standardized Mean Difference
Gender			
Male	45.8%	44.3%	.03
Female	54.2%	55.7%	-.03
Native American/American Indian	78.6%	79.4%	-.02
Free or reduced-price lunch eligibility	40.5%	40.5%	0
Special education designation	3.1%	3.1%	0
English language learner	1.5%	1.5%	0
Math pretest score*	-.31	-.33	.13
Grade			
4	11.5%	11.5%	0
5	16.8%	16.8%	0
6	20.2%	20.2%	0
7	9.2%	9.2%	0
8	29.4%	29.4%	0
10	13.0%	13.0%	0

*Variable on which baseline equivalence is established

Table F2. Background Characteristics of AVID-Core-Plus Students and Comparison Students at Baseline - Reading

Characteristic	AVID-Core-Plus Students (*n* = 256)	Comparison Students (*n* = 256)	Standardized Mean Difference
Gender			
Male	43.7%	44.5%	-.02
Female	56.3%	55.5%	.02
Native American/American Indian	80.5%	79.7%	.02
Free or reduced-price lunch eligibility	41.4%	42.2%	-.02
Special education designation	3.1%	3.9%	-.04
English language learner	1.6%	0.8%	.07
Reading pretest score*	-.37	-.33	.09
Grade			
4	11.3%	11.3%	0
5	16.8%	16.8%	0
6	21.1%	21.1%	0
7	9.4%	9.4%	0
8	29.3%	29.3%	0
10	12.1%	12.1%	0

*Variable on which baseline equivalence is established

Table F3. Background Characteristics of Schoolwide Students and Comparison Students at Baseline - Mathematics

Characteristic	Schoolwide Students (*n* = 356)	Comparison Students (*n* = 356)	Standardized Mean Difference
Gender			
Male	50.0%	52.2%	-.04
Female	50.0%	47.8%	.04
Native American/American Indian	82.0%	83.1%	-.03
Free or reduced-price lunch eligibility	47.2%	45.5%	.03
Special education designation	16.3%	17.4%	-.03
English language learner	1.1%	2.8%	-.12
Math pretest score*	-.91	-.78	-.01
Grade			
4	8.4%	8.4%	0
5	19.1%	19.1%	0
6	16.6%	16.6%	0
7	11.0%	11.0%	0
8	30.6%	30.6%	0
10	14.3%	14.3%	0

*Variable on which baseline equivalence is established

Table F4. Background Characteristics of Schoolwide Students and Comparison Students at Baseline - Reading

Characteristic	Schoolwide Students (*n* = 366)	Comparison Students (*n* = 366)	Standardized Mean Difference
Gender			
Male	50.8%	48.6%	.04
Female	49.2%	51.4%	-.04
Native American/American Indian	81.4%	78.7%	.07
Free or reduced-price lunch eligibility	45.4%	44.3%	.02
Special education designation	16.9%	14.8%	.06
English language learner	1.1%	1.1%	0
Reading pretest score*	-.88	-.81	.22
Grade			
4	8.2%	8.2%	0
5	17.8%	17.8%	0
6	16.7%	16.7%	0
7	10.1%	10.1%	0
8	31.7%	31.7%	0
10	15.6%	15.6%	0

*Variable on which baseline equivalence is established

College Readiness Outcomes

Table F5. Background Characteristics of AVID-Core-Plus Students and Comparison Students

Characteristic	AVID-Core-Plus Schoolwide Students (*n* = 32)	Comparison Students (*n* = 32)	Standardized Mean Difference
Gender			
Male	53.1%	53.1%	0
Female	46.9%	46.9%	0
Mean age (years)	17.1	17.3	-.22
Native American/American Indian	96.9%	100%	-.25
Parents' education level*	2.84	2.91	-.05
Grade level			
11th	56.3%	56.3%	0
12th	43.8%	43.8%	0

*Variable on which baseline equivalence is established

Table F6. Background Characteristics of Schoolwide Students and Comparison Students

Characteristic	Schoolwide Students (*n* = 31)	Comparison Students (*n* = 31)	Standardized Mean Difference
Gender			
Male	38.7%	54.8%	-.33
Female	61.3%	45.2%	.33
Mean age (years)	17.5	17.2	.35
Native American/American Indian	100%	100%	0
Parents' education level*	2.87	2.90	-.02
Grade level			
11th	58.1%	58.1%	0
12th	41.9%	41.9%	0

* Variable on which baseline equivalence is established

Appendix G: Analytic Results

In this appendix, we present specifications for our analytic models which predict relative academic achievement and college readiness for the AVID-Core-Plus sample and the schoolwide sample. The Analytic Methods section of Appendix A describes the analytic models used to obtain these estimates.

Tables G1 and G2 describe estimates for models predicting academic achievement. Table G1 describes estimates for the primary study, which predicts academic achievement among the AVID-Core-Plus treatment group and the comparison group. Table G2 describes results for the secondary study, which predicts academic achievement for the schoolwide group and the comparison group. Table G3 presents estimates predicting college readiness for the primary (AVID-Core-Plus sample) and secondary (schoolwide sample) analyses.

In the primary analyses predicting academic achievement among AVID-Core-Plus students and comparison group students, we find that being in the treatment group is not associated with any gains in reading or math standardized test scores relative to the comparison group. In these analytic models, we include relevant demographic characteristics, the cohort in the program, and the pretest. The baseline estimates reflect the treatment effect when all other variables' values in the model are held at zero. In order to facilitate interpretation of the baseline estimates, we have transformed some covariates so that they are mean-centered at zero. This means that the baseline estimate reflects the average student in our sample. The mean centering does not affect the treatment coefficient. We find that special education status and English language learner status is associated with decreases in relative math and reading test scores, and Cohort 1 is more likely to have increases in their relative math and reading test scores.

In our secondary analyses (Table G2), we observe decreases in relative test scores for both math and reading among students in the schoolwide intervention group versus the comparison group. Similar to results for the primary study, special education status and English language learner status were associated with decreases in test scores.

Table G3 describes estimates for analytic models predicting college readiness. In the primary study, in which we predict college readiness among AVID-Core-Plus students and comparison students, we find that being in the treatment group is associated with a significantly higher score on the CRI. In the secondary study, in which we predict college readiness among schoolwide intervention students and comparison students, we find no difference in college readiness between the treatment and comparison groups.

Table G1. Regression Model Specifications Predicting Academic Achievement, Primary Study (AVID-Core-Plus Sample)[71]

	Math		Reading	
	β	SE	β	SE
Time	-0.05	0.07	-0.11	0.07
Treatment Effect	-0.21	0.14	0.17	0.11
Race (Reference = American Indian)				
Not American Indian	-0.30	0.22	-0.01	0.20
Special education	-0.63**	0.23	-0.44*	0.09
English language learner	-0.63~	0.33	-0.79*	0.24
Gender (Reference = Male)	0.03	0.08	0.09	0.07
School-level pretest	0.81***	0.17	0.86***	0.12
Free/reduced price lunch eligibility	-0.08	0.12	-0.11	0.08
Cohort (Reference = Cohort 6)				
Cohort 1	0.40*	0.17	0.41*	0.12
Cohort 2	-0.10	0.16	0.00	0.11
Cohort 3	-0.05	0.16	0.13	0.11
Cohort 4	0.41	0.26	0.27	0.23
Cohort 5	-0.04	0.13	0.12	0.11
N	262		256	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ~ $p < 0.10$

[71] Pretest scores and indicator variables for gender, English Language Learner status, special education status, and free/reduced price lunch eligibility were mean-centered prior to inclusion in this model.

Table G2. Regression Model Specifications Predicting Academic Achievement, Secondary Study (Schoolwide Sample) [72]

	Math		Reading	
	β	**SE**	**β**	**SE**
Time	0.12~	0.07	0.08~	0.05
Treatment Effect	-0.27*	0.11	-0.25***	0.07
Race (Reference = American Indian)				
Not American Indian	0.06	0.20	0.20	0.17
Special education	-0.59***	0.09	-0.62***	0.08
English language learner	-0.05	0.24	-0.24	0.29
Gender (Reference = Male)	0.02	0.07	-0.01	0.06
School average test score	1.03***	0.12	0.73***	0.11
Free/reduced price lunch eligibility	-0.05	0.08	-0.04	0.07
Cohort (Reference = Cohort 6)				
Cohort 1	0.07	0.12	0.00	0.10
Cohort 2	0.07	0.11	0.16	0.11
Cohort 3	0.09	0.11	0.13	0.11
Cohort 4	0.06	0.23	-0.14	0.20
Cohort 5	0.17	0.11	0.12	0.10
N	356		366	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ~ $p < 0.10$

[72] Pretest scores and indicator variables for gender, English Language Learner status, special education status, and free/reduced price lunch eligibility were mean-centered prior to inclusion in this model.

Table G3. Regression Model Specifications Predicting College Readiness for the AVID-Core-Plus and Schoolwide Samples[73]

	AVID Core Plus Schoolwide		Schoolwide	
	β	**SE**	**β**	**SE**
Treatment Effect	3.10***	0.69	-0.15	0.67
Race (Reference = American Indian)				
Not American Indian[74]	-3.86	2.91	-	-
Age	-0.31	0.58	-0.02	0.52
Grade	1.67~	0.99	2.45**	0.83
Gender (Reference = Male)	0.78	0.70	3.21***	0.73
Parent education	0.16	0.27	0.43	0.28
Lives with siblings	-0.37~	0.20	0.10	0.18
Lives with parents	0.05	0.82	-0.75	0.89
Household resources[75]	0.87*	0.34	0.98***	0.27
N	64		62	

*** p <0 .001, ** p < 0.01, * p < 0.05, ~ p < 0.10

[73] Pretest scores and indicator variables for gender, English Language Learner status, special education status, and free/reduced price lunch eligibility were mean-centered prior to inclusion in this model.
[74] Age, gender, grade, household resources, parent education, siblings, and living situation were mean-centered prior to inclusion in this model.
[75] Respondents are asked on the CRQ to indicate if they have the following in their homes: a computer, a dictionary, a calculator, and more than 25 books. This variable was operationalized as an index from 0-4, indicating the number of affirmative responses to these four questions. This variable was then mean-centered prior to inclusion in the model.