LEVERAGING EXPERIMENTAL AND OBSERVATIONAL EVIDENCE TO ASSESS THE GENERALIZABILITY OF THE EFFECTS OF EARLY COLLEGES IN NORTH CAROLINA

Sarah Fuller
The University of North Carolina
Department of Public Policy
Education Policy Initiative at Carolina
Chapel Hill, NC 27599
sarah.fuller@unc.edu

Douglas Lee Lauen
(corresponding author)
The University of North Carolina
Department of Public Policy
Education Policy Initiative at Carolina
Chapel Hill, NC 27599
dlauen@unc.edu

Fatih Unlu
Rand Corporation
Santa Monica, CA 90401
fatihunlu@gmail.com

Abstract

Early college high schools (ECHSs) in North Carolina are small public schools of choice on college campuses that seek to promote attaining postsecondary credits in high school, college readiness, and postsecondary enrollment for underrepresented groups. Evidence from randomized control trials (RCTs) has shown positive effects of the ECHS model on important high school and postsecondary outcomes but appear to be underpowered to detect moderation effects. Furthermore, RCTs rarely address the key question of primary policy interest: Is the program effective on average across the population? This leaves us uncertain about (1) whether the early college intervention is a good strategy for helping to close enrollment and attainment gaps between under- and overrepresented groups, and (2) whether the expansion of the ECHS model will lead to the positive results that the RCT studies suggest. This study uses administrative data on all ECHSs in North Carolina including those that were part of a lottery study. This allows us to generate RCT estimates for the ECHSs in the lottery sample and quasi-experimental estimates for both the lottery and non-lottery ECHSs. We leverage this unique circumstance to generate estimates of the effect of ECHS on postsecondary outcomes that simultaneously maximize both internal and external validity. Specifically, because generalization depends on both moderation and sample selection, we (1) investigate sample selection, (2) conduct a moderation analysis to determine whether the effects of the intervention vary by key factors that also predict sample selection, and (3) produce a pooled estimate by extending a method called cross-design synthesis to incorporate both RCT evidence and quasi-experimental evidence. We find strong evidence that the positive results of the RCTs generalize to the full sample of ECHSs, which provides stronger evidence of effectiveness.

https://doi.org/10.1162/edfp_a_00379

1. INTRODUCTION

Early college high schools (ECHSs) are designed to provide students the opportunity to earn college credit by the time they graduate from high school. North Carolina was an early and intense adopter of this approach. In this state, ECHSs are small public schools of choice on college campuses that seek to promote college enrollment and success for underrepresented groups, such as poor and minority students. The rapid expansion of ECHSs across the country appears to represent an ideal model of evidence-based policy making. The effects of ECHSs have been evaluated through randomized control trials (RCTs) using data from lotteries used to assign students to overenrolled ECHS campuses (Berger et al. 2013, 2014; Edmunds et al. 2017, 2019; Song et al. 2021). These studies show positive effects of the ECHS model on important high school outcomes including attendance, being on track for college, high school graduation, and college credits accumulated while in high school (Berger et al. 2013, 2014) as well as postsecondary outcomes including attainment of postsecondary credentials (Edmunds et al. 2019; Song et al. 2021). Encouraged by this evidence, policy makers continue to invest in the model and new ECHSs have opened across more than thirty states (Webb 2014).

While the expansion of a model shown to be effective in high-quality studies is promising news, there are also important limitations of these RCTs regarding how broadly their results might apply, primarily because they used convenience samples of oversubscribed schools. An RCT on a convenience sample of schools that are both overenrolled and willing to use a random lottery to admit students provides strong evidence that the ECHS intervention can be effective, but it does not tell us whether the intervention works equally well in all settings or for all groups of students. In short, these RCTs are critical for assessing efficacy, but do not fully address the key question of primary policy interest: Is the program effective on average across the population? (Stuart et al. 2011). In technical terms, they do not answer the question of *generalizability*, that is, whether ECHSs are effective with types of students or types of communities that are not included in the set of sites included in the RCT. In order to know whether a scale-up of the model will show the same positive outcomes as those in the RCT study, we need to know whether the effects generalize to the whole population of current ECHSs and, further, whether they will generalize to students and communities that do not yet have an ECHS.

Even to the extent that the RCTs include a diverse group of schools and students, the small size of the studies limits their ability to shine light on heterogeneous effects of the intervention. Many individual-level RCTs are underpowered to detect moderation effects. Indeed, the literature reports mixed findings about whether the intervention has had larger effects for target populations—low-income students, disadvantaged minority students, and first-generation college students (Edmunds et al. 2017, 2019; Song et al. 2021). An existing quasi-experimental (QE) study[1] using a propensity score–matching design, on the other hand, produces compelling, but potentially biased, evidence of moderation effects (Lauen et al. 2017). This leaves us uncertain about (1) whether the

1. A note on terminology. We use the term randomized control trial (RCT) to refer to the design that exploits random assignment to evaluate an intervention with a treatment–control contrast. We use the term quasi-experiment (QE) to refer to a nonexperimental design of an intervention involving a treatment–comparison contrast, in which the comparison group is formed from nontreated cases that have equivalent baseline variable averages to the treatment group prior to the start of the intervention.

intervention is a good strategy for helping to close enrollment and attainment gaps between under- and overrepresented groups, and (2) whether the expansion of the ECHS model will lead to the positive results that the RCT studies suggest.

To resolve the tension between the internally valid but less generalizable results from relatively smaller RCTs and the arguably more generalizable[2] but less internally valid QE study with a large sample, this study combines evidence from both research designs to create a pooled estimate of the impacts of ECHSs that takes into account generalizability bias. Specifically, this paper has three aims:

1. to provide policy relevant information on the effects of ECHS on postsecondary outcomes across the whole population of ECHSs in North Carolina—a state with at least one ECHS in most of its 100 counties,
2. to provide estimates of the effects of ECHS on postsecondary outcomes for different student subgroups, and
3. to provide a methodological approach for combining an RCT study with population-level administrative data to produce more generalizable estimates, even in the presence of heterogeneous effects.

To achieve these aims, we use data on all ECHSs in North Carolina enrolling students between 2005–06 through 2010–11 that include those ECHSs that were part of an ongoing lottery study. This allows us to estimate RCT estimates for the ECHSs in the lottery sample and QE estimates for both the lottery sample and all other ECHSs in the state—the non-lottery sample (i.e., the ECHSs that did not participate in the RCT study in that particular year). We leverage this unique circumstance to generate estimates of the effect of ECHS on postsecondary outcomes that simultaneously maximize internal validity (the strength of the small RCT) and external validity (the strength of the large QE). Because generalization depends on both moderation and sample selection (Stuart et al. 2011), the present study:

1. investigates sample selection based on observable characteristics of the RCT sample,
2. conducts a moderation analysis using the QE to determine whether the effects of the intervention vary by key factors that also predict sample selection, and
3. produces a pooled estimate using a method called cross-design synthesis (Kaizar 2011) that incorporates both RCT and QE evidence.

Like prior research, our results show positive and significant impacts of ECHS attendance on postsecondary-related outcomes, including ACT scores, college courses completed during high school, four-year college enrollment, and two- and four-year college degree receipt. Our estimates of the impact on four-year college enrollment are somewhat larger than other studies. We also find heterogeneity in the impacts on

2. Some argue that internal validity is a precondition for generalizability; therefore, an estimate with low internal validity from a large sample may have no more generalizability than an RCT estimate from a nonrepresentative sample. It is precisely for this reason that our project uses a method (a cross-design synthesis) that is arguably less sensitive to confounding bias than standard techniques.

ECHS attendance with minority students and economically disadvantaged students benefitting more in four-year college enrollment and graduation than their more advantaged peers. Students who were low performing in eighth grade, however, do not benefit as much as their higher performing peers. Finally, and consistent with Lauen et al. (2017), ECHSs on four-year campuses have significantly different impacts than ECHSs on two-year campuses, with students being less likely to attend and graduate from two-year colleges but more likely to attend and graduate from four-year colleges. The present study differs from Lauen et al. (2017) in many important respects in that it: (1) focuses on postsecondary outcomes only, (2) includes more cohorts and more recent cohorts, (3) covers all types of postsecondary institutions through NSC data (public and private, in- and out-of-state), and (4) covers four-year degree completion. In addition, the present study is the first study of early colleges to combine the RCT evidence from the Edmunds et al. (2017) study with QE evidence to produce a pooled estimate.

The remainder of this paper proceeds as follows: First, we outline the early college intervention and prior research; second, we describe the data used for our three interrelated analyses; third, we describe the analysis methods and our results; and finally, we discuss the implications of our findings for research and policy.

2. BACKGROUND ON EARLY COLLEGE INTERVENTION AND EFFECTS

Early college high schools in North Carolina seek to re-envision the relationship between high school and college. They are small schools of choice (typically between one hundred and four hundred students), housed on mostly community college campuses, but also on some four-year campuses, intentionally designed and structured to provide all their students with an untracked, college-oriented environment, including early access to college courses. ECHSs start in ninth grade and students are expected to graduate within four to five years with a high school diploma and an associate degree or two years of college credit. ECHSs are designed to smooth the transition from high school to college for underrepresented student populations (Roderick et al. 2011). As such, the intervention primarily targets first-generation college students and students at risk of dropping out of high school. Nationally, 280 ECHSs in 31 states were funded and supported by the Bill and Melinda Gates Foundations Early College High School Initiative, which launched in 2001 (Edmunds et al. 2017). North Carolina, with its strong community college and state university systems, is currently home to over one hundred ECHSs, more than any other state. Each fall, roughly twenty thousand students spread across three quarters of the counties of North Carolina are educated in early colleges.

Edmunds et al. (2013) posit that a key to the success of ECHSs is "mandated engagement," in which the facilitators of student engagement such as peer attitudes, supervision, mentoring, parental involvement, and the like are so concentrated that student engagement with school is all but mandated. As new and small schools of choice designed around a shared mission, ECHSs recruit staff and students who believe in the mission and design principles. ECHSs raise academic rigor by enrolling students in college-level courses starting in freshman year. To help students meet these higher expectations, ECHSs are staffed with teachers, counselors, and administrators who understand that personalization and academic support are critical for student and organizational success.

Like magnet or charter schools, students choose whether to apply to an ECHS. Generally, students may only apply to the early college in their home county, although some early colleges accept students from multiple counties. Some ECHSs are oversubscribed, but many have slots for all who apply. In oversubscribed schools, lotteries are often, but not always, used to select which of the applicants will be invited to enroll. However, many ECHSs conduct screening interviews with families, and due to the rigorous nature of the curriculum at ECHSs, schools may also seek to recruit students who are interested in and academically prepared to complete a college-prep course of study. These two arguably conflicting aims—to serve underrepresented youth *and* students prepared to succeed in college-level coursework—combined with the fact that these are schools of choice, provides a strong justification for conducting moderation analyses to examine differences in impact across student populations (Lauen et al. 2017).

Studies using RCT and QE designs have reported positive impacts of ECHSs on a host of high school and postsecondary outcomes, including: small to moderate positive effects on English I end-of-course test scores and ACT scores; large positive effects on college credits accumulated while in high school and meeting grade-level targets for the courses necessary to gain admission to college; small positive or null effects on high school graduation; and large positive effects on postsecondary enrollment and postsecondary credential attainment, driven primarily by two-year institutions (Edmunds et al. 2012, 2013, 2017, 2019; Berger et al. 2013, 2014; Haxton et al. 2016; Lauen et al. 2017; Song et al. 2021).

While the effects for students overall are quite clear and positive, investigation of effects for subgroups (especially students underrepresented in higher education) has thus far produced much less conclusive evidence. An RCT study based in North Carolina reports positive differential effects (i.e., larger gains) on postsecondary enrollment for low-income and first-generation students compared with their counterparts, but negative differential effects (i.e., smaller gains) for college credits accumulated in high school and attainment of postsecondary credentials for underrepresented minorities and underprepared students (Edmunds et al. 2017, 2019). Not all differences in subgroups effects are statistically significant, including some differences as large as 8 percentage points, which suggests that this RCT study may be underpowered for probing heterogeneity in impacts. In the only nationwide study of ECHSs, a long-term study conducted by American Institutes for Research identified 154 ECHSs open by fall 2007, and subsequently discovered that two thirds were undersubscribed, so they could not hold a lottery. In the end only ten sites in five states (including two in the Edmunds and associates North Carolina sample) retained their lottery records, agreed to be a part of the study, and met other inclusion criteria. The most recent publication from the study (Song et al. 2021) reported that the impacts on postsecondary enrollment and degree attainment did not significantly differ between the subgroups analyzed, though, again the study was likely underpowered to detect moderation effects.

A QE study using a propensity score–matching design based on a larger sample of all early colleges in North Carolina reports positive ECHS impacts for both black and white students on college enrollment and completion, but to varying degrees. It reports that the ECHS impacts on four-year college enrollment for black students exceed those of white students, whereas the ECHS impacts on associate degree completion for white students exceed those of black students (Lauen et al. 2017). In addition, the

study notes that some ECHSs are hosted by four-year historically black college and university campuses rather than community colleges and that these sites have particularly large impacts on the enrollment of black students in four-year campuses after students graduate from high school.

In summary, prior research has established that the ECHS model can have positive impacts on student outcomes, including postsecondary enrollment and degree attainment. However, much of this evidence comes from a limited set of ECHSs that participated in lottery studies. Thus, the strength of the evidence on moderation and heterogeneous effects is limited. The existing larger QE study has more power to examine heterogeneous effects and is more generalizable to the full population of ECHSs at least within North Carolina, but due to the observational nature of the study, it has weaker internal validity than the RCT studies.

The present study provides a clearer answer to the question of whether positive impacts of postsecondary outcomes generalize across subgroups and to the broader range of ECHSs not in the lottery studies. It also uses an approach called cross-design synthesis to leverage the strengths of the RCT and the QE study to create an estimate that is both internally valid and generalizable.

3. DATA

Sample

The study team constructed a longitudinal dataset of student-level administrative data provided by the North Carolina Department of Public Instruction (NCDPI), the University of North Carolina (UNC) System, the North Carolina Community College System (NCCC), and the National Student Clearinghouse (NSC). Data provided by NCDPI cover the population of students who attended any public K–12 school in North Carolina during the 2004–05 to 2015–16 school years. The full dataset includes a total of more than 470,000 students in the six cohorts across the entire state. The RCT sample includes 19 ECHSs with at least one cohort in the RCT study, with an original randomized sample of 3,758 students. The research team implemented lotteries to divide students who applied to one of the study early colleges into two groups: those who were admitted (2,174 treatment group students) and those denied admission (1,584 control group students). Most control students ended up enrolling in regular high schools in their district. For some schools, lotteries were stratified to meet school administrators' priorities for admitting specific subgroups at higher rates (e.g., low-income and underrepresented minorities); therefore, some lotteries placed more students into the treatment group than the control group. The analyses accounted for the resulting unequal probabilities of treatment assignment using weights.[3] Overall, the compliance rate is

3. To better achieve the policy goal of recruiting underrepresented students into the intervention, most early colleges in the RCT sample identified priority populations (e.g., first-generation college attendees) for their incoming cohorts (fourteen of nineteen sites in the RCT sample). To include these schools in the sample and analyses, the research team stratified the eligible pool of applicants by the priority characteristics. Students from the priority groups had a better chance of being admitted through these stratified lotteries, which led to unequal probabilities of treatment assignment within the study sample. We included weights to account for the unequal probability of selection. If unaccounted for, such differences in treatment assignment probabilities would lead to baseline imbalance between the treatment and control groups. Each observation's weight was proportional to the inverse of the probability of its assignment to its respective groups. Weights were proportional to $1/P(T=1|X)$ and control weights were proportional to $1/(1-P(T=1|X)$, where $P(T=1|X)$ represents

91 percent, with 3.5 percent crossovers (students who were assigned to the control group but ended up enrolling in the ECHS to which they applied) and 12.9 percent no-shows (students who were assigned to the treatment group but did not enroll in any ECHS). Due to the retrospective nature of the study, we could not collect systematic data on oversubscription and implementation fidelity. As of the mid 2010s, many ECHSs were oversubscribed, recruited and selected students in very similar ways, and adherence to the early college design principles was supported by North Carolina New Schools, an intermediary organization charged with providing startup and ongoing technical assistance for the intervention on behalf of state policy makers.

We track individual students through administrative data continuously through high school and until they leave the public school system in North Carolina. At the student level, the data include demographic variables, test scores, middle school course taking, mobility, and educational classifications (details below).[4] Data provided by the UNC System and the NCCC System track student enrollments and graduations in North Carolina public two- and four-year colleges from 2009–10 to 2015–16. Data provided by the NSC tracks student enrollments and graduation in public and private colleges nationwide for students who graduated from NC public high schools from 2009 to 2015. The NSC data follows students from the time that they graduate high school until the spring of 2018 and covers postsecondary enrollments in the entire United States.[5] For college enrollment and graduation outcomes, students who do not appear in any of the college outcomes datasets are treated as not enrolled and as nongraduates.

The analytic sample for the present study consists of six cohorts of high school students who entered ninth grade for the first time in the 2005–06 through 2010–11 school years. To be included in these cohorts, students must have been enrolled in North Carolina public schools in eighth and then ninth grade in consecutive school years. This sample restriction is necessary to ensure that students have pretreatment demographic and performance data. We also limit the sample to high school graduates. This sample restriction is necessary for all college-going outcomes because only high school graduates were submitted to the NSC for matching, and we restrict the analytical sample across all outcomes to maintain a consistent sample across outcomes.[6]

the treatment assignment probability conditional on the covariate vector *X*. Weights were calibrated to ensure that the weighted sample size equaled the original sample size.

4. Overall, relatively few students are missing outcome data. Missingness for the outcomes ranges from 0 percent missing for college enrollment and graduation outcomes to 6.8 percent of students missing ACT scores. Somewhat larger percentages of students are missing one or more covariates. Very few students (less than 1 percent) are missing demographic information, but up to 15 percent are missing some prior test scores. However, missing covariates are imputed for these students using the "dummy variable" method, which entailed (i) replacing missing values for a given covariate with the sample mean and (ii) including an indicator for the imputed records in the propensity score and impact estimation models (Stuart 2010).
5. Dynarski, Hemelt, and Hyman (2015) estimate that NSC data covers greater than 95 percent of all colleges in North Carolina. In data received from the NSC, the number of Family Educational Rights and Privacy Act (FERPA) blocks of student records is approximately 1.3 percent of all enrollment records with the highest percentage of FERPA blocks concentrated in a small number of private colleges, including Duke University and Saint Augustine's. We acknowledge that missing records for some students is a limitation of the NSC dataset, but we have no reason to think that these minor coverage issues would invalidate the findings of this study.
6. Across the three cohorts approximately 15 percent of students who appear in ninth grade do not appear in eighth grade in the prior year and approximately 10 percent of eighth graders do not appear in ninth grade in the subsequent year. These excluded students consist primarily of those who were not enrolled in North Carolina public schools during one of the two years; students who were retained in either eighth or ninth grade; and a small number of students who could not be matched across time based on name, birthdate, and

The sample limitation to high school graduates could potentially bias our results if ECHS attendance has a significant impact on high school graduation. As shown in table A1, available in a separate online appendix that can be accessed on *Education Finance and Policy*'s Web site at https://doi.org/10.1162/edfp_a_00379, the estimated effect of ECHS on high school attendance is quite small. In addition, results indicate an increase in high school graduation at ECHSs in the QE design sample. Given that students on the margin of graduating are likely to be less academically prepared than other graduates, including these students in our outcome sample would be expected to bias us toward finding reduced college attendance among ECHS students.

Measures

Our study tracks six measures of college and career readiness. Two high school outcomes measure college readiness: ACT composite test score and number of college courses taken while enrolled in high school. The ACT is a mandated state test administered to all eleventh graders beginning in the 2011–12 school year. The number of college courses taken while in high school is the count of the total number of college credit bearing courses taken by a given student. This includes International Baccalaureate (IB) courses, Advanced Placement (AP) courses, and dual enrollment courses taken through two- or four-year colleges.[7] The other four study outcomes are measures of postsecondary enrollment and completion: two-year college enrollment, associate's degree completion, four-year college/university enrollment, and bachelor's degree completion. We do not combine postsecondary institution types because there are different economic and social returns to associate and bachelor's degrees (Jaeger and Page 1996; Dadgar and Trimble 2015). Two-year and four-year college enrollment are measured in the year immediately following high school graduation; associate's degree completion in the fourth year after high school graduation; and bachelor's degree completion in both the fourth and fifth year after high school graduation, in acknowledgement that many students do not complete their bachelor's degree program in four years. A summary of the outcome measures is shown in table 1.

This set of measures represents important academic and engagement measures for high school students and potential predictors of longer-term outcomes such as employment and wages, providing a more comprehensive view on the effectiveness of the intervention rather than high school or postsecondary outcomes alone. We note that measuring time to completion in this study is complicated by the fact that more than 96 percent of public high school students who obtain a high school diploma do so in four years while ECHS students finish in four or five years, depending on varying school policies and student preferences. In our sample, 42.4 percent of ECHS and 3.7 percent of non-ECHS graduates take five years to graduate from high school. Measuring bachelor's degree completion within five years of graduating from high

other administrative identification variables. The observations dropped due to grade repetition are duplicate observations across assigned to two cohorts. To fix this, we assign a student to the ninth-grade cohort where they first enrolled in ninth grade and drop them from the sample for the subsequent ninth-grade cohort where they would appear in ninth grade for the second time.

7. Two of the outcome measures—college course enrollment during high school and associate's degree receipt—are particularly closely tied to the ECHS intervention. However, we do not view them as "overaligned" with the treatment group because college course enrollment opportunities are available to students in the comparison conditions through dual credit and dual enrollment programs, AP programs, and IB programs.

Table 1. Outcome Measures Used in the Analysis

Measure	Description	Sample	Number of Cohorts and School Years
ACT	Composite of English, Reading, Math, and Science. Statewide sample statistics: mean = 18, standard deviation = 5.	All high school juniors	(2) 2009–10 through 2010–11
College credits in high school	Count of all college credit courses completed during high school, including AP and IB courses as well as dual enrollment courses.	All high school students in the state	(6) 2005–06 through 2010–11
2-year college enrollment	Enrolling in a two-year college in the year following graduating from high school.	High school graduates	(6) 2005–06 through 2010–11
Associate's degree completion (4th year)	Completing an associate's degree within four-years after graduating from high school.	High school graduates	(4) 2005–06 through 2008–09
4-year college/university enrollment	Enrolling in a four-year college or university in the year following graduating from high school.	High school graduates	(6) 2005–06 through 2010–11
Bachelor's degree completion (4th year)	Completing a bachelor's degree within four-years after graduating from high school.	High school graduates	(5) 2005–06 through 2009–10
Bachelor's degree completion (5th year)	Completing a bachelor's degree within five years after graduating from high school.	High school graduates	(4) 2005–06 through 2008–09

school gives ECHS students in five-year programs ten years from their high school freshman year to complete a degree, whereas students in non-ECHS get only nine years (four in high school and then five post high school). The alternative is to simply measure the years from freshman year and ignore the year in which a student had graduated from high school. This has the downside of losing comparability with postsecondary research, which generally tracks years beginning with the freshman year of college rather than the freshman year of high school. As it turns out, it makes little difference which approach is taken (results available from the authors upon request).

We also note that many studies measure degree completion for bachelor's degrees either six or eight years after beginning a degree program. Unfortunately, this is not possible with the time frame of data that we have available in this study, as the outcome data are only available four to five years after high school graduation for most cohorts. As we will show, the results related to bachelor's degree completion vary between whether completion was measured four or five years after high school graduation. It is our expectation that a measure taken six or eight years post high school graduation would show a weaker effect of ECHS enrollment on college completion. This would be consistent with the results of the two RCTs described above, both of which observed a decrease in impacts on four-year degree completion after four years post high school (Edmunds et al. 2019; Song et al. 2021).

The treatment variable in this study is an indicator for enrollment in an ECHS in ninth grade. This specification of the treatment variable does not include information on how many years a student remained enrolled in an ECHS. This choice of specification was made because the decision of a student to leave an ECHS may be related to their potential outcomes, as some types of students may find ECHS to be a better fit than other types of students. The analysis in this study also includes many student-level covariates that are measured during middle school, prior to enrolling in an ECHS. The covariates include student demographics, program participation, and performance. Demographics consist of race/ethnicity, sex, an indicator for being old for grade (measured

Table 2. Comparison of Early College High School (ECHS) Student Characteristics to Statewide Population

	Non-ECHS		ECHS	
	Mean	SD	Mean	SD
Asian	0.02	0.15	0.03	0.18
Black	0.28	0.45	0.27	0.44
Hispanic	0.07	0.26	0.10	0.29
American Indian	0.01	0.12	0.02	0.13
White	0.58	0.49	0.55	0.50
Multiracial	0.03	0.16	0.03	0.17
Economically disadvantaged	0.39	0.49	0.48	0.50
Limited English proficiency	0.04	0.20	0.04	0.20
Disability	0.10	0.30	0.04	0.21
Academically gifted	0.19	0.40	0.22	0.41
Mobility	0.17	0.38	0.22	0.41
Old for grade	0.14	0.35	0.10	0.30
Male	0.49	0.50	0.39	0.49
Mid sch read average	0.13	0.90	0.40	0.74
Mid sch math average	0.15	0.92	0.37	0.78
Science score eighth	0.01	0.69	0.21	0.71
Passed Algebra mid sch	0.22	0.41	0.28	0.45
Mid sch absences	6.45	5.39	6.44	5.16
Urban setting	0.38	0.48	0.34	0.47
Cohort1 (05–06)	0.16	0.36	0.05	0.21
Cohort2 (06–07)	0.16	0.37	0.13	0.33
Cohort3 (07–08)	0.16	0.37	0.15	0.36
Cohort4 (08–09)	0.17	0.38	0.20	0.40
Cohort5 (09–10)	0.17	0.38	0.23	0.42
Cohort6 (10–11)	0.17	0.38	0.24	0.43
Observations	466,634		13,171	

Notes: This table includes the means and standard deviations for all students enrolled in early college high schools and non-early college high schools (all other public high schools) statewide in the cohorts included in the study. See text for details on calculation of standardized bias. Mid sch = middle school.

as being fifteen years old on 1 September of the ninth-grade year), an economically disadvantaged indicator, and a mobility indicator (measured as having changed schools between sixth and eighth grade). Program participation variables include indicators for being identified with a disability, as academically and intellectually gifted, and as limited English proficiency. Performance covariates include middle school math test score average, middle school reading test score average, eighth-grade science test score, indicators for taking and passing algebra in middle school, and average middle school absences. Middle school average variables including absences and math and reading test scores are measured by taking the average of all measures during sixth to eighth grade.

Additional variables that are used as moderators in the heterogeneity analysis but not as covariates in regressions include an urban setting indicator and an indicator for an ECHS hosted on the campus of a four-year college. These variables are measured at the school level. Table 2 shows the means and standard deviations of all baseline measures for ECHS and non-ECHS schools.

4. ANALYTIC METHODS

Our goal is to estimate the population average treatment effect of attending an ECHS in North Carolina.[8] As this is a retrospective study and the RCT sites were not selected randomly from the statewide population, to establish a claim to generalizability we must assume unconfounded treatment selection and unconfounded sample selection (Imai et al. 2008; Stuart et al. 2011). The first assumption, *unconfounded treatment selection*, requires that treatment assignment is random and independent of potential outcomes given observed covariates, a reasonable assumption in a randomized design. For the QE estimates for the non-lottery sites, we rely on best practices within the QE literature and work our team conducted on a within-study comparison to reduce treatment selection bias (Unlu et al. 2021).

To maintain *unconfounded sample selection,* the second assumption, there must be no unmeasured variables that are related to both sample selection and the treatment effect across strata of the unmeasured variable. This requires that all potential moderators of the treatment effect that are also potential predictors of sample selection are measured and properly controlled for in the analysis. In short, to ensure generalizability with respect to one potential observable moderator, M (e.g., gender, race, ethnicity) one must have either (a) no sample selection based on M, or (b) no effect heterogeneity based on M. In short, if one can rule out either sample selection or heterogeneity with respect to M, then one can establish a strong claim to generalizability at least with respect to that one observable factor. For example, suppose that the proportion of female students in lottery ECHS sites is larger than the non-lottery sites or the state. This would only harm generalizability if there is also heterogeneity in the effect of the ECHS intervention by gender.

Our strategy for reducing sample selection error, and thus improving generalizability, relies on first assessing the nature and extent of the potential sample selection problem, exploring moderation, and creating a pooled estimate for both lottery and non-lottery sites using an approach called cross-design synthesis (Kaizar 2011; Pressler and Kaizar 2013). The resulting population effect estimate is the sum of the RCT impact from the lottery sites and a generalizability bias parameter that adjusts the RCT impact by its estimated external validity bias, as we explain below.[9] To create an unbiased generalizable pooled estimate, cross-design synthesis rests on the assumption that the size of any bias present in the QE estimate relative to the true effect of the intervention is the same across the lottery and non-lottery samples. We use moderation analyses to assess the validity of this assumption with respect to observables and use a novel weighted cross-design synthesis approach to adjust for potential concerns related to heterogeneity in bias across subgroups.

8. We define the population as all ECHS in North Carolina. A key reason for this choice is due to the strong influence of an intermediary organization (North Carolina New Schools), which provided start-up support, technical assistance, professional development, and a common set of design principles for all ECHSs across the state. It is quite possible that ECHSs share many similarities with ECHSs in other states, but we cannot be as sure about the fidelity to the North Carolina New Schools approach to ECHSs in other states as we can in North Carolina. It is of value to researchers and policy makers, however, to know if North Carolina's particular model of ECHS implementation is broadly successful across the whole state.
9. Stated differently, the cross-design synthesis estimate is an average of the RCT impact estimate for the lottery sites and the QE effect estimate for the non-lottery sites, where the latter is adjusted for the treatment selection error estimated using the within-study comparison.

Representativeness of the RCT

The ability of an RCT to generalize to the whole population of interest rests on the assumption that the RCT sample is representative of the population of interest. We begin our analysis by assessing the extent to which our lottery sample differs from our non-lottery sample. To do this, we present tables of standardized mean differences between the lottery and non-lottery ECHS samples for each fixed or pretreatment variable. We compute standardized bias for continuous variables as:

$$d = \frac{\left(\bar{x}_{lottery} - \bar{x}_{nonlottery}\right)}{\sqrt{\frac{s^2_{lottery} + s^2_{nonlottery}}{2}}},$$

where $\bar{x}$ represents the subsample mean for variable x and s^2 is the subsample variance of variable x. For dichotomous outcomes we define standardized bias as:

$$d = \frac{\left(\hat{p}_{lottery} - \hat{p}_{nonlottery}\right)}{\sqrt{\frac{\hat{p}_{lottery}\left(1-\hat{p}_{lattery}\right) + \hat{p}_{nonlottery}\left(1-\hat{p}_{nonlattery}\right)}{2}}},$$

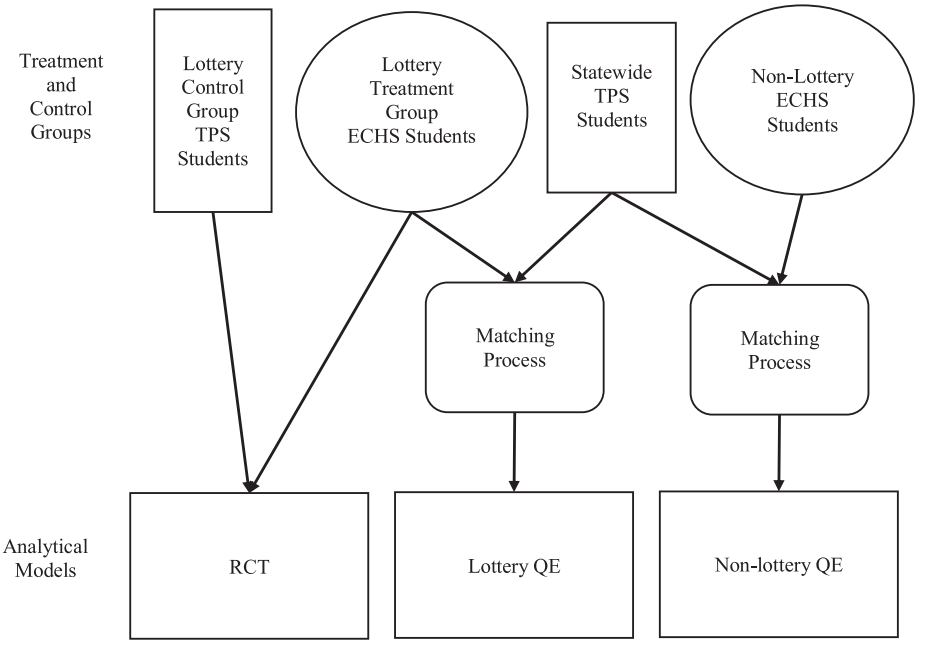
where $\hat{p}$ is the prevalence or mean of the dichotomous variable in the subsample (Austin 2009).

Unlike a t-test of statistical significance, standardized bias (d) is a measure of substantive significance. Fixing the mean difference, t will be larger when N is larger, will therefore reject the null even when mean differences are substantively small (e.g., below a tenth of a standard deviation in effect size). We also present standard t-tests with unequal variances assumed, however, because apparently large effect sizes can be statistically insignificant due to small sample size. Given a relatively large sample size, even some trivial differences will likely be statistically significant with a t-statistic $> |1.96|$. Covariates for which the lottery and non-lottery samples have a standardized bias greater than 0.10 will be identified as subgroups across which the lottery sample is not representative. Covariates on which the lottery sample is not representative will be the subgroups of greatest concern for our generalizability analysis (as described above), especially in cases where there are also detectable heterogeneous effects by subgroup.

ECHS Impact on Postsecondary Outcomes

The next step in our analysis is to calculate the impact of ECHS attendance on our high school and postsecondary outcomes. We divide ECHS schools into two samples according to whether they participated in the lottery RCT study. Together, these two samples—the lottery sample and the non-lottery sample—include all ECHSs in the state of North Carolina. Figure 1 provides an illustration of the connection between the treatment and comparison groups and the analytical models used to generate these three estimates.

For the *lottery ECHS*, we calculate the impact two different ways: (1) an experimental analysis, and (2) a QE estimate on the RCT sample. The first is the experimental estimate, which is calculated using the RCT sample and fitting a regression model that controls for the lottery indicators and student-level covariates listed above. The second estimate comes from a doubly robust QE method in which we fit a propensity-score model using the treatment students in the RCT sample and all non-ECHS students in North Carolina. We first fit a probit model with being selected into a lottery ECHS



Notes: TPS = traditional public school; ECHS = Early College High School.

Figure 1. Samples and Models

through the lottery in the RCT study as the dependent variable, and baseline (middle-school) covariates described in the previous section as predictors of enrolling in an ECHS. Using the estimated model, we next predict the probability of being a lottery ECHS student in ninth grade for each student in the sample (non-lottery ECHS students and lottery non-ECHS students are removed from the sample for this analysis). From these predicted probabilities, $\hat{p}$, we then compute inverse probability of treatment weights as 1 if student i was a lottery ECHS student and $\frac{\hat{p}}{(1-\hat{p})}$ if student i was not a lottery ECHS student. We examine weighted and unweighted balance statistics to ensure that weighting reduces baseline differences between the ECHS and non-ECHS students.[10] Finally, we fit an outcome regression weighted inverse to the probability of being in a lottery ECHS. In this regression, enrolling in a lottery ECHS indicator is the treatment indicator of interest. Baseline predictors of enrolling in an ECHS and cohort fixed effects are included as controls. We use ordinary least squares (OLS) models for continuous outcomes (ACT and college courses) and linear probability models for the binary postsecondary outcomes (two- and four-year college enrollment and graduation).[11] The outcome model controls for the same set of covariates used in the estimation of the propensity score to increase the precision of the impact estimates

10. We consider any covariate on which the standardized mean difference between the treatment and comparison group exceeds 0.1 to be out of balance. Online appendix tables A2 and A3 show that for both lottery and non-lottery samples, the unweighted comparison group is out of balance with the treatment group on many covariates, including the percent of economically disadvantaged students and prior test scores. However, once propensity-score weights are applied, the weighted comparison group is in balance with the treatment group on all covariates and standardized mean differences are very small—generally less than 0.01.
11. Logit model results for binary outcomes are included in online appendix table A4.

and for double-robustness (Bang and Robins 2005). Using the same baseline characteristics in the propensity model and outcome model gives the analyst two ways to get the "right" model specification (once in the propensity model and another time in the impact model for the outcome measure).

For the *non-lottery ECHS*, we use a similar doubly robust propensity-score inverse probability of treatment weighted outcome regression as a QE method. These analyses yield three estimates of attending ECHS for each outcome: the RCT or experimental estimate for lottery ECHS, the QE estimate for lottery ECHS, and the QE estimate for non-lottery schools. Difference between the RCT and QE estimates for the lottery ECHS—different estimation approaches on the same sample—reveal bias in the QE estimate because of the exclusion of unmeasured covariates which predict ECHS enrollment and outcomes (i.e., confounders). Differences between the QE estimate for the non-lottery ECHS and the RCT estimates for the lottery ECHS—different estimation approaches and different samples—may be driven by either bias in the QE estimate or by real differences in the effectiveness of the lottery and non-lottery ECHS. In later steps of the analysis, we will use the estimate of bias generated by comparing the two estimates on the RCT sample to quantify the real differences in the effectiveness of the lottery and non-lottery ECHS and generate a pooled estimate using cross-design synthesis.

Heterogeneous Effects across Subgroups

Having generated estimates of the effects of ECHS across the whole sample, our next step is to explore heterogeneity across subgroups within the sample. Exploratory analysis, the target populations of the intervention, and prior research guided our search for potential moderators. The target subpopulations covered in prior research have been gender, underrepresented minority (black, Hispanic, Native American, multiracial), economically disadvantaged, low-performing, and first-generation students. Data limitations preclude the present study from performing a moderation analysis on first-generation students because there is no variable for parental education for most of our cohorts. We include the other target populations in the present study. In addition, we examine two potential site-level factors that prior research has shown to be particularly relevant for postsecondary enrollment: urbanicity and type of campus host (two- or four-year campus). Distance to a university is inversely related to the probability of attending a postsecondary institution (Alm and Winters 2009). Four-year universities in North Carolina are concentrated in urban areas. Therefore, we would expect that students who attend a rural ECHS would be less likely to enroll in a four-year college or university than students who attended an urban ECHS. Prior work has also shown a strong tendency of campus host type (two-year versus four-year) to influence type of enrollment in postsecondary institutions post high school and completion of an associate degree versus bachelor's degree (Lauen et al. 2017). Given these findings about moderation, it is important to include these site-level factors in a moderation and sample selection analysis.

We use the non-lottery sample for moderation analysis because the sample sizes of the lottery sample are quite small for many of the subgroups. For each *focal* subgroup and its *counterpart* subgroup (e.g., for male as the focal subgroup, female is the counterpart subgroup), we conduct doubly robust propensity-score inverse probability

of treatment weighted outcome regressions. These regressions follow the QE method described for the lottery and non-lottery ECHS samples. However, for subgroup analyses, the generation of propensity scores, weighting, and regression analysis are run separately for each subgroup of interest and for its counterpart (e.g., underrepresented minority and not an underrepresented minority). This results in two estimates of the ECHS effect, one for the subgroup and one for its counterpart. We assess the size of the difference and the statistical significance using a z-test (Altman and Bland 2003):

$$Z = \frac{\beta_{g1} - \beta_{g2}}{\sqrt{SE\left(\beta_{g1}\right)^2 + SE\left(\beta_{g2}\right)^2}}.$$

Where we find a statistically significant difference in the estimated effect of the ECHS intervention between a subgroup of interest and its counterpart, we will consider the effect of the ECHS to be heterogeneous. Subgroups with heterogeneous effects will be identified as subgroups of concern for our generalizability analysis if the lottery sample is also unrepresentative for these subgroups.

Pooled Impact Estimate for Generalization: The CDS Estimator

Having estimated separate models of ECHS impact for the lottery ECHS and non-lottery ECHS samples, we now turn to creating a pooled estimate of the average impact of ECHS enrollment for all ECHS in the state. While it is possible to create a combined QE estimate that includes all ECHSs in the state, this estimate would have potential bias resulting from the exclusion of unmeasured covariates. In many studies, this potentially biased estimate is the best available strategy, but in this study, we are in the unique position of having an RCT estimate for a subsample of the population of interest, which can be used to validate the QE model and estimate the size of the bias. This unique situation allows us to increase the generalizability of our estimate.

We use a technique ideally suited to the scenario, in which the same intervention has been implemented both experimentally and nonexperimentally and the same outcome data are available for all sites in the population of interest. This approach, akin to meta-analysis, is called *cross-design synthesis* (CDS). First proposed by the General Accounting Office for medical effectiveness research and formalized by Kaizar (2011) and Pressler and Kaizar (2013), this technique stratifies the observational study units by inclusion in the RCT sample to estimate the sample selection bias of the RCT. In theory, one would want an RCT estimate from both the sample of sites included in the RCT population and the sample of sites not included in the RCT population. If this were possible, one could compute a sample-size weighted average of two stratified randomized estimators to obtain the PATE (Kaizar 2011):

$$\widehat{PATE} = \frac{N_{X=1}}{N} D^R\left(X = 1\right) + \frac{N_{X=0}}{N} D^R\left(X = 0\right),$$

where $X = 1$ means the unit is in population stratum included in the RCT, $X = 0$ means the unit is the population stratum excluded from the RCT, and D^R is the impact from an RCT (rather than an observational study). But, because it is impossible to calculate an RCT impact for units excluded from the RCT strata, $D^R(X = 0)$, we estimate it by adjusting the experimental estimate for the effect of the inclusion criteria as estimated from the observational data:

$$\hat{D}^R(X=0) = D^R(X=1) + \left[D^O(X=0) - D^O(X=1)\right],$$

where D^O is the impact from the QE. That is, the experimental (i.e., unbiased) effect for the excluded sample is set to the adjusted version of the experimental effect for the included sample where the adjustment is equal to the difference of the QE estimate between the excluded and included samples. Plugging in the estimate $\hat{D}^R(X=0)$ for $D^R(X=0)$ into the PATE equation above produces the cross design synthesis estimate for the entire population (Kaizar 2011):

$$\widehat{PATE}_{CDS} = D^R(X=1) + \frac{N_{X=0}}{N}\left[D^O(X=0) - D^O(X=1)\right],$$

or, the RCT impact on the sites in the lottery study and a generalizability bias parameter, weighted by the fraction of the sample not in the lottery study. Because this technique involves a difference in the QE estimate for the included sample and the QE estimate for the excluded sample, the CDS estimator is unbiased when the QE estimates' treatment selection error (or selection bias) is constant across the two samples.[12]

The CDS estimates the true effect of the intervention across the whole population by adjusting the RCT estimate for sample selection bias. It does this by quantifying sample selection bias using the difference between the QE estimate for the RCT sample and the QE estimate for the rest of the population, excluding the RCT sample with the assumption that differences in the impact estimate across these two samples are driven by differences in sample since the estimation approach is the same.

Pooled Impact Estimates that Account for Heterogeneity: The Extended CDS Estimator

The primary assumption of the CDS allows for the QE estimates to be biased but requires that the size of the bias be uniform between the lottery ECHS and the non-lottery ECHS. In North Carolina, the presence of a strong intermediary in North Carolina New Schools assured that all ECHSs across the state were following similar models for the intervention. Under this condition, if the lottery and non-lottery ECHSs were in similar communities and served similar groups of students, the assumption of similar bias in the QE estimates for the two samples might be plausible. However, significant differences in the makeup of the student body and the locales of the lottery and non-lottery schools make this assumption less plausible. This is especially true if there is also heterogeneity in the effect of the ECHS intervention across these student or school subgroups.

Based on the subgroup heterogeneity analysis combined with the representativeness of the lottery sample analysis, we identify subgroups as potentially concerning with regard to the ability of the lottery estimates and the CDS to give us an unbiased pooled estimate, because we have evidence that suggests both the composition of the lottery and non-lottery samples and the effect of the intervention were different for these subgroups. We extend the CDS estimator to incorporate the problem of unrepresentativeness of the RCT sample and heterogeneity in effects between subgroups. We do this by estimating the components of the CDS estimator separately for each subgroup and

12. For a proof of the bias and other properties of the CDS estimator consult appendix B of Kaizar (2011).

using a weighted combination across subgroups to generate an overall CDS estimate. The resulting extended CDS estimator is:

$$\widehat{PATE}_{CDS} = \frac{N_{W=1}}{N}\left\{D^R\left(X=1, W=1\right) + \frac{N_{W=1,X=0}}{N_{W=1}}\left[D^O\left(X=0, W=1\right) - D^O\left(X=1, W=1\right)\right]\right\} + \frac{N_{W=0}}{N}\left\{D^R\left(X=1, W=0\right) + \frac{N_{W=0,X=0}}{N_{W=0}}\left[D^O\left(X=0, W=0\right) - D^O\left(X=1, W=0\right)\right]\right\}.$$

In the equation above, X, D^R, and D^O are defined in the previous section and W denotes membership in a subgroup for which the impact varies. Without loss of generality, we assume two subgroups:[13] $W=1$ for membership in the focal subgroup (e.g., female students) and $W=0$ denotes its counterpart (e.g., male students). N_W and N_X are the number of individuals in each group defined by values of X and W. This Extended CDS estimator uses the information from the RCT and the QE on the lottery-based sample to separately adjust for bias in the QE for the non-lottery sample for each subgroup, and then recombines the separate estimates into a pooled estimate by weighting according to the percent of the population in each subgroup. We will compare the pooled estimates derived from this subgroup weighted CDS to the pooled estimates from the overall CDS with no subgroup information. We note that this extension of the CDS estimator requires that the subgroup and its counterpart both be represented in the lottery sample. If a relevant subgroup is not represented in a stratum, it is not possible to calculate all components of the CDS for that group.

This CDS estimator is unbiased when:

(1) $E[\Delta_T^O(X=1, W=1)] = E[\Delta_T^O(X=0, W=1)]$ and
(2) $E[\Delta_T^O(X=1, W=0)] = E[\Delta_T^O(X=0, W=0)]$,

where $\Delta_T^O(X, W) = D^O(X, W) - D^R(X, W)$, which is the bias (i.e., treatment selection error) of the observational estimator for given X and W. These conditions hold if either C1 or C2 holds:

C1:

(a) $T \perp U \mid X, W$ (U is the unobserved confounder for the QE estimator) or
(b) $E[Y \mid T, X, W, U] = g(T, X, W)$ for some function g

C2:

(a) $X \perp U \mid T, W$ and $X \perp U \mid W$ and
(b) $E[Y \mid T, X, W, U] = g1(T, X, W) + g2(T, U, W)$ for some functions $g1$ and $g2$.

13. To be clear, the Extended CDS estimator is not limited to only one subgroup, but in practice the curse of dimensionality limits the number of subgroups that can be reliably incorporated into a stratification-based estimator.

Table 3. Comparison of Students Enrolled in Lottery and Non-Lottery Early College High School (ECHS)

	Lottery			Non-Lottery			Diff. Statistics	
	M	SD	*N*	M	SD	*N*	SB	*t*
Asian	0.01	0.11	2,159	0.04	0.19	10,986	−0.17	−8.57
Black	0.28	0.45	2,159	0.27	0.44	10,986	0.03	1.29
Hispanic	0.08	0.27	2,159	0.10	0.30	10,986	−0.07	−3.13
Native American	0.00	0.06	2,159	0.02	0.14	10,986	−0.14	−8.10
White	0.59	0.49	2,159	0.55	0.50	10,986	0.09	1.64
Multiracial	0.04	0.19	2,159	0.03	0.17	10,986	0.04	3.71
Economically disadvantaged	0.52	0.50	2,155	0.47	0.50	10,986	0.09	3.65
Limited English proficiency	0.04	0.18	2,155	0.04	0.20	10,986	−0.03	−1.45
Disability	0.04	0.21	2,155	0.05	0.21	10,986	−0.01	−0.41
Academically gifted	0.21	0.41	2,155	0.22	0.41	10,986	−0.02	−0.69
School mobility	0.23	0.42	2,132	0.22	0.41	10,986	0.03	2.22
Old for grade	0.12	0.32	2,174	0.10	0.30	10,986	0.05	1.34
Male	0.40	0.49	2,159	0.39	0.49	10,986	0.02	1.03
Mid sch read average	0.31	0.75	2,163	0.41	0.74	10,986	−0.13	−5.43
Mid sch math average	0.26	0.76	2,165	0.38	0.79	10,986	−0.16	−6.92
Gr 8 Science score	0.21	0.85	1,784	0.20	0.70	10,986	0.01	0.49
Passed Algebra Mid sch	0.23	0.42	2,114	0.28	0.45	10,986	−0.13	−5.65
Urban campus	0.28	0.45	2,168	0.34	0.47	10,985	−0.13	−5.67
Four-year campus	0.18	0.39	2,174	0.12	0.33	10,986	0.16	6.35

Notes: Students are considered lottery students if they are enrolled in one of the nineteen ECHSs that participated in the RCT study in a cohort for which a lottery was held. All other ECHS students in the relevant cohorts are counted as non-lottery students. See text for details on how we calculate standardized bias (SB). *t*-statistic is *t* with unequal variances assumed. Mid sch = middle school.

In the equations above, U is an unobserved variable that is correlated with the treatment assignment in the observational study (i.e., treatment-outcome confounder). The conditions discussed in this section apply when there are multiple unobserved confounders. T is an indicator for the treatment group, that is, $T = 1$ for those in the treatment group and $T = 0$ for those in the control group. Condition C1 suggests there is no treatment confounding within levels of X and W. C1(a) and C1(b) capture the trivial cases in which the observational estimator is unbiased, therefore CDS is unbiased. In these cases, the conventional CDS is also unbiased unless W eliminates any residual confounding conditional on X only. C2(a) indicates that confounding due to sample inclusion (X) and other variables (W are U) are separate and the average outcome does not depend on any interaction between X are U conditional on W. This condition implies that the bias of the observational estimator is the same across levels of X for a given value of W. This is a weaker condition that the assumption of the conventional CDS estimator that implies that the bias of the observational estimator is the same across levels of X, which may be less plausible if the treatment impact varies by W.

5. RESULTS

Representativeness of the RCT

The first step in our generalization exercise is to assess the representativeness of the RCT sample compared to excluded members of the population of interest, in this case the non-lottery sample. Table 3 displays the means and standard deviations for all

Table 4. Randomized Control Trial (RCT) and Quasi-experimental (QE) Impact Estimates for Lottery and Non-Lottery Early College High School (ECHS)

	ACT Score	College Courses in HS	2-Year Enrollment	2-Year Grad	4-Year Enrollment	4-Year Grad 4th Year	4-Year Grad 5th Year
Lottery RCT Estimates							
ECHS	0.284^{+}	11.451***	−0.093***	0.237***	0.084***	0.079***	0.055^{+}
	(0.167)	(0.519)	(0.023)	(0.031)	(0.023)	(0.020)	(0.032)
Control mean	19.304	2.344	0.371	0.169	0.380	0.144	0.232
N	1,912	3,265	3,269	2,088	3,269	2,088	1,034
Lottery QED Estimates							
ECHS	0.394**	11.88***	−0.0292	0.325***	0.0435*	0.0515**	0.00604
	(0.145)	(0.793)	(0.0186)	(0.0439)	(0.0201)	(0.0167)	(0.0212)
Comp. Mean	18.33	1.546	0.310	0.0764	0.381	0.164	0.262
N	150,672	462,276	463,177	378,960	463,177	378,960	298,410
Non-Lottery QED Estimates							
ECHS	0.435***	11.30***	−0.0229*	0.282***	0.0363**	0.0780***	0.0416**
	(0.113)	(0.560)	(0.00979)	(0.0301)	(0.0135)	(0.0112)	(0.0150)
Comp. Mean	19.78	2.082	0.292	0.0780	0.441	0.194	0.293
N	155,152	476,543	477,620	390,168	477,620	390,168	307,470

Notes: This table displays three sets of impact estimates: Lottery RCT, Lottery Sample QE, and Non-Lottery Sample QE. For each set of estimates, the table displays the coefficient, standard error, adjusted control/comparison mean, and number of observations. For continuous outcomes, the coefficient is estimated using a linear regression and can be interpreted as the change in predicted outcome from a switch from a non-ECHS to an ECHS. For binary outcomes, the coefficient is estimated using a linear probability model, which can be interpreted as the percentage point change in the likelihood of an outcome from a switch from a non-ECHS to an ECHS. Treatment and control margins can be interpreted as the expected average outcome for the treatment group if they attended an ECHS or a traditional public school, respectively. QED = quasi-experimental design. $^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

pretreatment covariates for the lottery and non-lottery samples as well as the standardized bias for each variable. Several variables show imbalance across the lottery and non-lottery samples. The Asian and American Indian ethnic groups are both imbalanced across the two samples; however, these groups are extremely small, so small differences in the actual number of students from the two groups result in significant differences between the samples. The average middle school math and reading test scores of students in the non-lottery sample are higher than in the lottery sample by about a tenth of a standard deviation. In addition, students in the non-lottery sample are more likely to pass algebra in middle school. These imbalances suggest that ECHS students in the non-lottery sample are higher performing on average than those in the lottery ECHS sample. In addition, there are imbalances in the number of rural versus urban schools and the number of schools on four-year college campuses. The non-lottery sample is more likely to be urban but less likely to be on a four-year college campus. These imbalances between the lottery and non-lottery samples indicate that the potential for sample selection error will be important to consider in our heterogeneity and generalization analysis.

ECHS Impact on Postsecondary Outcomes

The second step of our analysis is to generate three sets of estimates of the effect of ECHS attendance on our college-going outcomes of interest. Table 4 shows RCT and QE impacts for the lottery and non-lottery samples. In the first two columns, we present OLS regression coefficients. For the postsecondary outcomes we present linear

probability coefficients (OLS estimated with a binary outcome).[14] To contextualize the size of the reported impacts, we include control and comparison margins, which are the adjusted predictions for those who did not enroll in an ECHS. For ACT scores, all three models show positive impacts, although the estimate for the lottery RCT is smallest and only marginally significant. The estimates are all quite small, less than half a point on a 32-point scale. However, relative to the size of the estimates, the differences between the estimates are considerable with the QE estimate for the lottery sample being more than a third larger than the lottery RCT estimate, suggesting considerable treatment selection bias, and the non-lottery QE estimate 10 percent larger than the lottery QE estimate, suggesting more moderate sample selection bias.

In contrast, the estimates of the effect on college courses taken in high school are quite similar across the three models and the effect size is very larger. Although control and comparison students have opportunities to take college courses in high school through AP or IB courses and the state's dual enrollment program, on average, they only take 1.5 to 2.3 such courses, depending on the specific comparison group. On the other hand, ECHS students average more than 13 college courses while in high school, or the equivalent of nearly three semesters of college.

Considering two-year college outcomes, estimates are quite different across the three models. For two-year enrollment, all models show negative point estimates, but the size of the coefficient for the lottery RCT model is nearly three times as large as the other two estimates. This difference from the lottery QE model suggests that the QE model is upwardly biased. In substantive terms, the lottery RCT models show a large decrease, more than 25 percent, in students enrolling in two-year colleges after high school. However, this is driven by students who either completed their associate's degree during high school or went straight to a four-year college. In keeping with this, the effect of ECHS attendance on two-year college graduation is substantively large in all three models. The smallest estimate, from the lottery RCT model, suggests that two-year graduation rates more than doubled among students who attended an ECHS. This is particularly meaningful in policy terms because two-year graduation rate is measured four years after high school graduation, so even with 200 percent of the expected time to degree, control students are not catching up with ECHS students on this outcome. In the comparison across the estimates, the lottery QE estimate is 37 percent larger than the lottery RCT estimate, indicating substantial treatment selection bias. The lottery QE is also approximately 15 percent larger than the non-lottery QE, suggesting more moderate sample selection bias.

Turning to four-year college outcomes, all three estimates suggest positive impacts on four-year college enrollment, ranging from 3.6 percentage points to 8.4 percentage points. These effect sizes are moderate to large in policy terms. Unlike the two-year college outcomes, the lottery QE estimate appears to be downwardly biased relative to the lottery RCT. The lottery QE remains upwardly biased relative to the non-lottery QE. For four-year college graduation, the RCT estimate finds substantively large positive impacts, 7.9 and 5.5 percentage points. In contrast, the lottery QE estimate for

14. All binary outcome models have also been run as logistic models (see online appendix table A4). The results from the logistic models are substantively the same and similar in magnitude. We chose to focus on linear probability models (LPMs) for ease of interpretation of the magnitude of the effects.

Table 5. Combined Statewide Estimate of Early College High School (ECHS) Impact Using Cross-Design Synthesis

	RCT Estimate	Lottery ECHS Student *N*	Non-Lottery ECHS Student *N*	Proportion of Non-Lottery Students	QE Non-Lottery Estimate	QE Lottery Estimate	$PATE_{CDS}$
ACT composite	0.284	1,129	4,632	0.804	0.435	0.394	0.317
College courses in HS	11.451	1,903	10,971	0.852	11.300	11.880	10.957
2-year enrollment	−0.093	1,907	10,986	0.852	−0.023	−0.029	−0.088
2-year grad 4th year	0.237	1,129	7,463	0.869	0.282	0.325	0.200
4-year enrollment	0.084	1,907	10,986	0.852	0.036	0.044	0.078
4-year grad 4th year	0.079	1,129	7,463	0.869	0.078	0.052	0.102
4-year grad 5th year	0.055	511	5,269	0.912	0.042	0.006	0.088

Notes: Randomized control trial (RCT) and quasi-experimental (QE) estimates are ordinary least squares coefficients for linear outcomes and linear probability model coefficients for binary outcomes. $PATE_{CDS}$ is the population average treatment estimate from the cross-design synthesis equation shown on page 584.

four-year graduation five years after high school graduation is nonsignificant and close to zero. The lottery QE estimate is thus downwardly biased compared to the lottery RCT estimate. The lottery QE estimate also appears to be downwardly biased relative to the non-lottery QE estimates.

In substantive policy terms, the results thus far are in keeping with the prior literature with positive impacts on college courses in high school, two-year degree receipt, four-year enrollment, and four-year graduation. However, differences between the two sets of QE estimates suggest that the lottery sample may not be fully generalizable. Specifically, it may provide effect sizes for two-year enrollment, two-year graduation, and four-year graduation that are upwardly biased and estimates for four-year graduation that are downwardly biased.[15] However, differences between the two estimates on the lottery sample caution against simply using a QE to generate a more generalizable estimate, because the QE estimates appear to contain substantial bias related to unobservables. In the next section, we use the CDS estimator to create a pooled estimate that attempts to correct for these two sources of bias.

Pooled Impact Estimate for Generalization

Table 5 shows the CDS estimation for each outcome. The CDS estimate is calculated by adjusting the RCT estimate for sample selection bias by using the difference between the QE estimates on the two different samples as a measure of the bias. Thus, for example, the sample selection bias for two-year college graduation would be equal to −0.043 (0.325 − 0.282). Because non-lottery students make up 86.9 percent of the population for that outcome, the RCT estimate would be adjusted by −0.037 (−0.043 × 0.869), leading to a pooled estimate of 0.200.

Looking at our individual outcomes, the pooled estimate for ACT score is somewhat larger than the lottery estimate, while college course taking in high school is somewhat smaller. Two-year and four-year enrollment look very similar in the pooled estimate to the lottery impact. However, two-year degree receipt is smaller in the pooled estimate and four-year year degree receipt is larger. As these differences represent between 16

15. The differences in the direction of the bias may be due to differences in how representative the lottery sample is for the different cohorts included in the estimates for different outcomes.

and 60 percent of the effect size of the RCT estimate and roughly 3 percentage points in college graduation rates, the differences are substantively meaningful. These estimates account for sample selection bias but rely on the assumption that treatment selection bias is the same across the QE on the lottery sample and the QE on the non-lottery sample.

In addition to our analyses in the previous section assessing the representativeness of the lottery sample, we also assess this assumption of similar bias between the lottery and non-lottery samples by comparing the coefficients from the selection models used to generate propensity scores. These results are shown in online appendix table A5. The results show that treatment selection is similar for many observable characteristics, including identifying as black, economically disadvantaged, and middle school math test scores, but differs significantly for some other characteristics, including identification as academically and intellectually gifted and reading test scores. These differences are consistent with the patterns seen in table 3 related to the unrepresentativeness of the lottery sample on some characteristics. We believe that these differences emphasize the importance of directly considering sample selection error into the lottery sample and heterogeneity of impacts as we do with our extend CDS estimate, however, we acknowledge that we cannot fully eliminate the possibility of differences in bias from unobservables between the lottery and non-lottery samples.

Heterogeneous Effects across Subgroups

The next section of the results examines heterogeneity in the effect of ECHS attendance across subgroups of students and schools. These impacts may vary from previously reported literature from the sample of lottery sites because of sample differences such as urbanicity and campus host. In addition, power is higher to detect effects in this analysis due to a larger sample size. Table 6 shows the QE estimates of the impact of ECHS attendance for each subgroup and its counterpart across all ECHSs in the state of North Carolina (table 6 shows linear probability models for binary outcomes; logit results are in online appendix table A6). Our findings related to student subgroups indicate three consistent patterns: (1) the ECHS impact on underrepresented groups on ACT scores is equivalent to the ECHS impact on non-underrepresented groups, (2) the impact on number of college courses generally disfavors underrepresented groups (i.e., underrepresented minorities [URM] get a smaller increase in college courses compared to their non-underrepresented peers), and (3) the effects on underrepresented groups on postsecondary enrollment and completion favors these groups relative to their counterparts. While the differences in effect size for ACT and college courses are small, the differences for college enrollment and completion are often large. To take just one example, consider five-year graduation. The ECHS impact for URM students is .089, whereas for non-URM students it is .016. The difference in the subgroup impacts is .074 ($p < .05$).[16] The same pattern can be observed for economic disadvantaged students compared to non-economically disadvantaged students.

16. While the subgroup difference in four-year graduation rate between URM and non-URM students is not statistically significant at the .05 level in the linear probability model (LPM) results, it is in the logit model results. See online appendix table A4. The same pattern of nonsignificance in LPM and significance in logit results for some outcomes is also observed in the subgroup differences between low- and high-performing students and economically disadvantaged and non-economically disadvantaged students.

Table 6. Subgroup Effect Estimates Using the Non-Lottery Quasi-experimental Analysis

Underrepresented Minority			Not an Underrepresented Minority		Subgroup Differences	
	ECHS	SE	ECHS	SE	Difference	sig ($p < .05$)
ACT score	0.564	(0.126)	0.446	(0.124)	0.118	
College courses in HS	10.00	(0.632)	12.17	(0.579)	−2.170	*
2-year enrollment	−0.020	(0.016)	−0.026	(0.009)	0.006	
2-year grad 4th year	0.217	(0.035)	0.322	(0.033)	−0.105	*
4-year enrollment	0.084	(0.021)	0.010	(0.012)	0.074	*
4-year grad 4th year	0.098	(0.016)	0.072	(0.012)	0.026	
4-year grad 5th year	0.089	(0.025)	0.016	(0.012)	0.074	*
Economically Disadvantaged			**Not Economically Disadvantaged**		**Subgroup Differences**	
	ECHS	SE	ECHS	SE	Difference	sig ($p < .05$)
ACT score	0.626	(0.116)	0.419	(0.129)	0.207	
College courses in HS	10.70	(0.564)	11.99	(0.572)	−1.290	
2-year enrollment	−0.038	(0.012)	−0.023	(0.010)	−0.015	
2-year grad 4th year	0.235	(0.028)	0.315	(0.036)	−0.080	
4-year enrollment	0.082	(0.015)	0.006	(0.013)	0.076	*
4-year grad 4th year	0.097	(0.011)	0.080	(0.014)	0.017	
4-year grad 5th year	0.089	(0.016)	0.018	(0.016)	0.071	*
Male			**Female**		**Subgroup Differences**	
	ECHS	SE	ECHS	SE	Difference	sig ($p < .05$)
ACT score	0.480	(0.132)	0.404	(0.114)	0.076	
College courses in HS	10.92	(0.577)	11.57	(0.590)	−0.650	
2-year enrollment	−0.017	(0.011)	−0.031	(0.011)	0.017	
2-year grad 4th year	0.253	(0.030)	0.300	(0.032)	−0.047	
4-year enrollment	0.027	(0.016)	0.043	(0.013)	−0.016	
4-year grad 4th year	0.081	(0.014)	0.078	(0.012)	0.003	
4-year grad 5th year	0.049	(0.016)	0.038	(0.016)	0.011	
Low Performing in Eighth Grade			**Not Low Performing in Eighth Grade**		**Subgroup Differences**	
	ECHS	SE	ECHS	SE	Difference	sig ($p < .05$)
ACT score	0.649	(0.118)	0.490	(0.114)	0.159	
College courses in HS	7.884	(0.544)	12.27	(0.576)	−4.386	*
2-year enrollment	−0.022	(0.016)	−0.030	(0.009)	0.008	
2-year grad 4th year	0.136	(0.022)	0.319	(0.034)	−0.183	*
4-year enrollment	0.092	(0.019)	0.026	(0.012)	0.065	*
4-year grad 4th year	0.067	(0.012)	0.088	(0.012)	−0.021	
4-year grad 5th year	0.072	(0.017)	0.040	(0.016)	0.032	
Two-Year Hosted ECHS			**Four-Year Hosted ECHS**		**Subgroup Differences**	
	ECHS	SE	ECHS	SE	Difference	sig ($p < .05$)
ACT score	0.405	(0.111)	0.836	(0.306)	−0.431	
College courses in HS	11.77	(0.612)	8.022	(0.824)	3.748	*
2-year enrollment	−0.010	(0.008)	−0.103	(0.022)	0.093	*
2-year grad 4th year	0.341	(0.023)	−0.029	(0.003)	0.370	*
4-year enrollment	0.017	(0.011)	0.176	(0.028)	−0.159	*
4-year grad 4th year	0.062	(0.008)	0.167	(0.028)	−0.105	*
4-year grad 5th year	0.019	(0.009)	0.131	(0.036)	−0.112	*

Table 6. Continued.

	Rural		Urban		Subgroup Differences	
	ECHS	SE	ECHS	SE	Difference	sig ($p < .05$)
ACT score	0.650	(0.122)	0.186	(0.215)	0.464	
College courses in HS	12.45	(0.625)	9.160	(0.868)	3.290	*
2-year enrollment	−0.034	(0.008)	−0.008	(0.021)	−0.026	
2-year grad 4th year	0.346	(0.024)	0.169	(0.046)	0.177	*
4-year enrollment	0.032	(0.012)	0.050	(0.025)	−0.018	
4-year grad 4th year	0.076	(0.009)	0.083	(0.022)	−0.007	
4-year grad 5th year	0.038	(0.010)	0.040	(0.027)	−0.002	

Notes: For each subgroup and its counterpart, the table displays the coefficient and standard error (SE). For continuous outcomes, the coefficient is estimated using a linear regression and can be interpreted as the change in predicted outcome from a switch from a non-ECHS (Early College High School) to an ECHS. For binary outcomes, the coefficient is estimated using a linear probability model, which can be interpreted as the percentage point change in the likelihood of an outcome from a switch from a non-ECHS to an ECHS. For each pair of subgroups, the table also displays the difference in the coefficients and a calculation of the statistical significance of the difference. sig ($p < .05$) column has a star if the subgroup means are significantly different at the $p < .05$ level.

In addition to student subgroups, we examine two school-level subgroups. Compared with ECHSs hosted on four-year college campuses, two-year hosted sites produce larger gains in the number of college courses taken in high school, smaller negative effects on two-year college enrollment after high school, and larger increases in two-year degree receipt. On the other hand, four-year hosted ECHSs produce larger impacts on four-year college outcomes than do two-year hosted sites.[17] Compared with urban ECHS, rural ECHS see greater increases in the number of college courses taken during high school and higher two-year graduation rates.

The results suggest that, contrary to prior research (Edmunds et al. 2017, 2019; Song et al. 2021), ECHSs are successfully meeting their goal of closing gaps in four-year college enrollment and graduation between the target populations and more advantaged peers. Minority, low-income, and low-performing students are less likely to take college courses in high school at an ECHS compared to classmates, but they see larger gains in four-year college enrollment and graduation. There are also notable differences between ECHS located on two-year campuses and those on four-year campuses when it comes to college enrollment and graduation, suggesting that the transition to college is very different depending on the campus location of the ECHS.

Our analysis to this point suggests that there is some reason to be concerned with a simple extrapolation of RCT impacts to the whole sample of ECHSs in North Carolina. Although the QE estimates across the two samples are only moderately different, there are several subgroup characteristics on which there appears to be sample selection into the lottery sample. Notably, low-performing, rural, and four-year campuses are overrepresented in the lottery sample. This is particularly concerning because we also find heterogeneity in effects for several subgroups, including minority students, low-income students, low-performing students, and students at ECHSs on four-year college campuses. We focus on subgroups for which there is evidence of both sample

17. We note that these findings are not tautological because our postsecondary outcomes are measured after students receive a high school diploma.

Table 7. Summary Table of All Estimates Using Cross-Design Synthesis (CDS)

	RCT Estimate	CDS without Subgroup Weights	Low-Performing/ High-Performing CDS	Two-year/ Four-year CDS	4 Category CDS
ACT composite	0.284	0.317	0.210	0.341	0.297
College courses in HS	11.451	10.957	10.977	11.736	11.660
2-year enrollment	−0.093	−0.088	−0.086	−0.091	−0.086
2-year grad 4th year	0.237	0.200	0.206	0.201	
4-year enrollment	0.084	0.078	0.079	0.099	0.092
4-year grad 4th year	0.079	0.102	0.100	0.125	
4-year grad 5th year	0.055	0.088	0.093		

Notes: Estimates are drawn from table 5 and online appendix tables A5 and A6. Estimates for last column not shown in online appendix, but available from authors upon request. RCT = randomized control trial.

selection and heterogeneity—low-performing students and students attending ECHS on four-year campuses—as particularly concerning for our pooled impacts.

Pooled Impact Estimates that Account for Heterogeneity

The simple CDS does not consider any potential differences in treatment selection bias that might be created by differences in the composition of the lottery and non-lottery samples. In this section, we calculate an extended CDS estimator that explicitly accounts for differences in bias between subgroups that are differentially represented within the lottery and non-lottery samples. The two sets of subgroups we focus on are low-performing versus high-performing students and ECHS on two-year college campuses versus those on four-year college campuses. Table 7 shows the results of these analyses compared to the original RCT estimates and the simple CDS estimate (see online appendix tables A7 and A8 for extended CDS estimates by student performance and campus location, respectively; the four-category CDS is available upon request from the authors). For ACT scores, which are on a 32-point scale, the RCT and CDS estimates do not vary substantively. The effect on the number of college courses taken in high school remains similar across all estimates but is slightly larger once we adjust for campus location. The effect on two-year enrollment remains negative and similar in magnitude in all estimates. The effect on two-year completion is positive in all cases, but about 15 percent smaller in the CDS columns. The effect on four-year enrollment is positive across estimates, but the estimated effect size is more than 25 percent (2 percentage points) larger in the CDS estimate, accounting for campus location. Four years after high school graduation, the effect size for the CDS estimate considering campus location is 23 percent (2.3 percentage points) larger than the simple CDS. These comparisons show the importance of considering differences in sample selection bias across subgroups, especially in the presence of large heterogeneous effects in the case of campus location and four-year college outcomes.

6. SUMMARY AND DISCUSSION

In summary, across the statewide set of ECHSs, the effect of ECHS attendance on student outcomes includes a small increase in ACT scores, a large increase in college course-taking in high school, a fairly large decrease in two-year enrollment, an increase in four-year college enrollment, a large increase in two-year college graduation, and a

smaller increase in four-year college graduation. These top-level findings are generally in line with prior RCT studies of ECHSs (Edmunds et al. 2019; Song et al. 2021). We find a somewhat larger impact on degree receipt at four years compared to other studies. Prior studies show a fadeout in the ECHS advantage in four-year degree receipt as students move from four years post-high school to six years post-high school. The current study cannot follow students to six years post-high school but shows somewhat larger impacts at four and five years post-high school, particularly for non-lottery schools. This suggests that existing RCT evidence may be understating the size of the effects of ECHSs on four-year completion due to biases in their sample of schools. Understanding the size of effects is particularly important for evaluating the costs and benefits of the ECHS program.

Our heterogeneity analysis shows that the impact of attending an ECHS varies across student subgroups. URM students experience smaller impacts than non-URM students on the number of college courses taken in high school, but larger impacts on four-year college outcomes. Economically disadvantaged students and low-performing students also experience larger impacts on four-year college outcomes. Low-performing students see smaller gains in high school course taking and two-year college graduation. The largest differences occur between students who attended ECHSs hosted on two-year campuses and those who attended ECHSs on four-year campuses. Students on four-year campuses take fewer college courses in high school and are less likely to enroll in or graduate from a two-year college, but they are much more likely to enroll in and graduate from a four-year college.

Compared with prior research, our subgroup findings are somewhat unique. Most prior research has not found differential impacts on four-year college degree receipt for minority or economically disadvantaged students (Edmunds et al. 2019; Song et al. 2021). This difference may be due to the larger sample size and greater power to detect small differences in this study. Our examination of four-year versus two-year college location is also unique and important given the large differences in effects. The analyses suggest that the representation of four-year campus-located ECHSs in lottery studies is particularly relevant for producing generalizable results.

It is notable that while the patterns of impact heterogeneity between the statewide analyses presented here and those reported for the lottery ECHS are generally consistent, heterogeneity in impact estimates is more pronounced when all ECHSs in North Carolina are examined. This is not surprising, given that analyses that compare impact estimates between two subgroups tend to have larger sample size requirements than analyses of average treatment effects.

We also found some differences between the effects yielded by the CDS approach that ignored impact heterogeneity and representativeness of the RCT sample and the CDS approach that took these factors into account. This suggests that to create a valid estimate using an approach such as the CDS when there is evidence for heterogeneity in impacts, it is necessary for the RCT to be representative of all subgroups or to be large enough to allow for separate analyses across the subgroups. Unfortunately, a stratification estimator such as the extended CDS is limited by the curse of dimensionality. One cannot reliably estimate quantities in sparsely populated or empty cells of a cross-classification.

Finally, as noted above, this paper utilized a rich statewide administrative dataset that covered twelve school years. This provided us with access to (i) multiple treatment cohorts (which may have played a role in the observed heterogeneity in impact estimates), (ii) a rich set of baseline covariates including measures of student achievement from multiple years, and (iii) outcome measures for treated students who participated in the lotteries conducted for the RCT as well as a much larger sample of treated students who were not included in the RCT analyses. The latter made the CDS approach possible. We acknowledge that this may not be feasible for all RCTs conducted with purposive or convenience samples. Fortunately, there is a growing research base that offers others approaches, such as stratification and weighting, that can be utilized to probe the generalizability of impact estimates in such cases (Stuart et al. 2011, 2015; Olsen et al. 2013; Tipton 2014a, 2014b). Another promising approach is to combine lottery and non-lottery effects within an empirical Bayes framework to compensate for the inability of many RCT studies to conduct heterogeneous effects of interventions due to limited power (Bruhn 2020). A promising avenue for future research would be to compare the performance of the CDS approach to these alternative approaches.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305A150477 to the University of North Carolina at Chapel Hill. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors gratefully acknowledge the support of the North Carolina Department of Public Instruction, the University of North Carolina System, and the North Carolina Community College System. Special thanks go to project advisors Tom Cook, Julie Edmunds, and Elizabeth Stuart; research assistants Elc Estrera, Joshua Horvath, and Anna Rybinska; and policy advisors Alisa Chapman, Andrew Kelly, Bill Schneider, and Dan Cohen-Vogel. All errors and opinions belong to the authors. Authors listed alphabetically.

REFERENCES

Alm, James, and John V. Winters. 2009. Distance and intrastate college student migration. *Economics of Education Review* 28(1): 728–738. 10.1016/j.econedurev.2009.06.008

Altman, Douglas G., and J. Martin Bland. 2003. Interaction revisited: The difference between two estimates. *BMJ* 326(7382): 219. 10.1136/bmj.326.7382.219

Austin, Peter C. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* 28(25): 3083–3107. 10.1002/sim.3697

Bang, Heejung, and James M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4): 962–973. 10.1111/j.1541-0420.2005.00377.x

Berger, Andrea, Lori Turk-Bicakci, Michael Garet, Joel Knudson, and Gur Hoshen. 2014. *Early college, continued success: Early College High School Initiative impact study*. Washington, DC: American Institutes for Research. Available https://www.air.org/sites/default/files/AIR_ECHSI_Impact_Study_Report-_NSC_Update_01-14-14.pdf.

Berger, Andrea, Lori Turk-Bicakci, Michael Garet, Mengli Song, Joel Knudson, Clarisse Haxton, Kristina Zeiser, Gur Hoshen, Jennifer Ford, Jennifer Stephan, Kaeli Keating, and Lauren Cassidy.

2013. *Early college, early success: Early College High School Initiative impact study*. Washington, DC: American Institutes for Research. Available https://eric.ed.gov/?id=ED577243.

Bruhn, Jesse. 2020. The consequences of sorting for understanding school quality. Working paper. Available https://sites.google.com/site/jessebruhn3/jesse_bruhn_jmp.pdf? Accessed 16 June 2021.

Dadgar, Mina, and Madeline Joy Trimble. 2015. Labor market returns to sub-baccalaureate credentials: How much does a community college degree or certificate pay? *Educational Evaluation and Policy Analysis* 37(4): 399–418. 10.3102/0162373714553814

Dynarski, S. M., S. W. Hemelt, and J. M. Hyman. 2015. The missing manual: Using National Student Clearinghouse data to track postsecondary outcomes. *Educational Evaluation and Policy Analysis* 37(1 suppl): 53S–79S. 10.3102/0162373715576078

Edmunds, Julie A., Lawrence Bernstein, Fatih Unlu, Elizabeth Glennie, John Willse, Arthur Smith, and Nina Arshavsky. 2012. Expanding the start of the college pipeline: Ninth-grade findings from an experimental study of the impact of the early college high school model. *Journal of Research on Educational Effectiveness* 5(2): 136–159. 10.1080/19345747.2012.656182

Edmunds, Julie A., Fatih Unlu, Jane Furey, Elizabeth Glennie, and Nina Arshavsky. 2019. What happens when you combine high school and college? The impact of the early college model on postsecondary performance and completion. Working paper. Available https://serve.uncg.edu/uploads/hsreform/Postsecondary_performance_and_completion_Merging_high_school_rev_10_16_19_identified.pdf.

Edmunds, Julie A., Fatih Unlu, Elizabeth Glennie, Lawrence Bernstein, Lily Fesler, Jane Furey, and Nina Arshavsky. 2017. Smoothing the transition to postsecondary education: The impact of the early college model. *Journal of Research on Educational Effectiveness* 10(2): 297–325. 10.1080/19345747.2016.1191574

Edmunds, Julie A., John Willse, Nina Arshavsky, and Andrew Dallas. 2013. Mandated engagement: The impact of early college high schools. *Teachers College Record* 115(7): 31. 10.1177/016146811311500705

Haxton, Clarisse, Mengli Song, Kristina Zeiser, Andrea Berger, Lori Turk-Bicakci, Michael Garet, Joel Knudson, and Gur Hoshen. 2016. Longitudinal findings from the early college high school initiative impact study. *Educational Evaluation and Policy Analysis* 38(2): 410–430. 10.3102/0162373716642861

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2): 481–502. 10.1111/j.1467-985X.2007.00527.x

Jaeger, David A., and Marianne E. Page. 1996. Degrees matter: New evidence on sheepskin effects in the returns to education. *Review of Economics and Statistics* 78(4): 733–740. 10.2307/2109960

Kaizar, Eloise E. 2011. Estimating treatment effect via simple cross design synthesis. *Statistics in Medicine* 30(25): 2986–3009. 10.1002/sim.4339

Lauen, Douglas L., Nathan Barrett, Sarah Fuller, and Ludmila Janda. 2017. Early colleges at scale: Impacts on secondary and postsecondary outcomes. *American Journal of Education* 123(4): 523–551. 10.1086/692664

Olsen, Robert, Larry Orr, Stephen Bell, and Elizabeth A. Stuart. 2013. External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management* 32(1): 107–121. 10.1002/pam.21660

Pressler, Taylor R., and Eloise E. Kaizar. 2013. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Statistics in Medicine* 32(20): 3552–3568. 10.1002/sim.5802

Roderick, Melissa, Vanessa Coca, and Jenny Nagaoka. 2011. Potholes on the road to college: High school effects in shaping urban students' participation in college application, four-year college enrollment, and college match. *Sociology of Education* 84(3): 178–211. 10.1177/0038040711411280

Song, Mengli, Kristina Zeiser, Drew Atchison, and Iliana Brodziak de los Reyes. 2021. Early college, continued success: Longer-term impact of early college high schools. *Journal of Research on Educational Effectiveness* 14(1): 116–142. 10.1080/19345747.2020.1862374

Stuart, Elizabeth A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25(1): 1–21. 10.1214/09-STS313

Stuart, Elizabeth A., Catherine P. Bradshaw, and Philip J. Leaf. 2015. Assessing the generalizability of randomized trial results to target populations. *Prevention Science* 16(3): 475–485. PMCID: 4359056. 10.1007/s11121-014-0513-z

Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2): 369–386. 10.1111/j.1467-985X.2010.00673.x

Tipton, Elizabeth. 2014a. How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics* 39(6): 478–501. 10.3102/1076998614558486

Tipton, Elizabeth. 2014b. Stratified sampling using cluster analysis: A balanced-sampling strategy for improved generalizations from experiments. *Evaluation Review* 37(2): 109–139. 10.1177/0193841X13516324

Unlu, Fatih, Douglas Lee Lauen, Sarah Crittenden Fuller, Tiffany Berglund, and Elc Estrera. 2021. Can quasi-experimental evaluations that rely on state longitudinal data systems replicate experimental results? *Journal of Policy Analysis and Management* 40(2): 572–613. 10.1002/pam.22295

Webb, Michael. 2014. *Early college expansion: Propelling students to postsecondary success, at a school near you*. Boston, MA: Jobs for the Future.