Methods for Policy Analysis

Burt Barnow, Editor

CAN QUASI-EXPERIMENTAL EVALUATIONS THAT RELY ON STATE LONGITUDINAL DATA SYSTEMS REPLICATE EXPERIMENTAL RESULTS?

Fatih Unlu, Douglas Lee Lauen, Sarah Crittenden Fuller, Tiffany Berglund, and Elc Estrera

Abstract

Do quasi-experimental (QE) studies conducted with baseline covariates that are typically available in the longitudinal administrative state databases yield unbiased effect estimates? This paper conducts a within-study comparison (WSC) study that compares experimental impacts of early college high school (ECHS) attendance with QE impacts drawn from the state and locales. We find that (1) QE models for outcomes with natural (matching) pretests replicated the randomized benchmarks quite well; (2) the replication bias is not sensitive to type of propensity score model or method; and (3) imposing locational restrictions (i.e., local matching) on the comparison students—specifically choosing them from among non-treatment students who came from the same feeder middle schools as the treatment students—does not decrease the QE bias; on the contrary, it performed worse than the models that did not impose this restriction for most outcomes. The first two findings are generally consistent with other education WSCs while the third one is not, suggesting that in cases where selection may be driven by individual-level factors, such as this one, local matching may yield biased treatment effect estimates by greatly reducing the pool of potential comparison units and distorting balance on unobservable confounders while prioritizing balance on observable factors.

INTRODUCTION

Randomized controlled trials (RCTs) are considered the strongest research design for estimating causal impacts of programs. They result in statistically equivalent groups and well designed and implemented RCTs yield unbiased impact estimates in expectation. Despite their popularity among empirical researchers, RCTs are not always ethical, feasible, or cost-effective. The primary challenge to conducting RCTs is to get program implementers, potential participants, and other stakeholders to agree to randomization, which requires all parties to give up control over who does

Journal of Policy Analysis and Management, Vol. 40, No. 2, 572–613 (2021)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.22295

and does not get access to the intervention. Feasibility challenges associated with RCTs often limit external validity and statistical power because conducting an RCT is sometimes possible only with a small and selective sample of volunteers who agree to random assignment. Additional analyses or design features are needed to assess the generalizability of results beyond such a study sample (Tipton & Olsen, 2018). In addition, most RCTs are conducted prospectively; therefore, the cost and time required to design and implement all study components (e.g., creating recruitment and data collection protocols and data analysis plans, recruitment of participants, conducting random assignment, monitoring the integrity of random assignment, collecting data, and conducting analysis) can preclude conducting large scale and longitudinal RCTs.

Fortunately, we are living in a time in which the digital revolution has produced a great deal of administrative data. This "data tsunami" (Decker, 2014) makes both RCTs and quasi-experimental (QE) studies easier to conduct. QE approaches involve an intervention that precedes measurement of an outcome, but with nonrandom selection of treatment and comparison groups. These types of designs can be prospective or retrospective. In addition, they tend to face fewer feasibility challenges and often have more statistical power than RCTs.

While the theoretical underpinnings of the conditions that lead to biased QE results are generally well understood (Shadish, Cook, & Campbell, 2002), there is still a significant gap in the research base about the practical aspects of internal validity concerns associated with the use of QE methods. Emerging first in the job training literature and then spreading to other fields including education, design replication studies or within study comparisons (WSCs) aim to fill this gap and inform the designs and analyses of QE studies by comparing plausibly unbiased impact estimates from an RCT to multiple QE effect estimates for the same intervention using the same data and measures (Dehejia & Wahba, 1999; Franker & Maynard, 1987; Heckman, Ichimura & Todd, 1997, 1998; LaLonde, 1986; Smith & Todd, 2005). In short, WSCs assess the correspondence between the treatment-control contrast from an RCT with a treatment-comparison contrast from a QE. In most WSCs (known as "dependent arm" WSCs; Wong & Steiner, 2018), the two contrasts involve the same treatment group. The only difference is that the QE arm includes a nonexperimentally generated comparison group in the place of the experimentally generated control group. These studies empirically investigate whether it is possible to replicate results of RCTs using QE methods, the magnitude and direction of bias in QE estimates of program effects, and the specific design features or analysis methods that minimize bias in QE designs and support causal inferences.

WSC research in education is rapidly developing but there are still some important gaps in the existing knowledge base (Wong, Valentine, & Miller-Bain, 2017). One outstanding question pertains to whether researchers should expect to obtain accurate (i.e., unbiased) effect estimates from QE studies conducted with the baseline covariates typically available in the state longitudinal data systems (SLDSs). The existing WSCs in education highlight that the pretreatment version of the outcome (pretest) is the most important covariate for minimizing QE bias (Cook & Steiner, 2010; Dong & Lipsey, 2018; Hallberg et al., 2018; Wong et al., 2017). But some educational outcomes do not have natural pretests because they are one-time events. This specifically applies to many important outcome measures that are examined by interventions targeting high school and postsecondary students, such as high school graduation, being academically prepared for college, and college enrollment, persistence, and graduation. Therefore, whether QE analyses for these types of outcomes obtained from extant administrative data could produce valid effect estimates remains an important and unanswered question. A related question is concerned with whether augmenting the extant set of covariates with geographical restrictions on the set of comparison units influences the magnitude of QE bias. Acknowledging

the wide variety of QE analysis options available to researchers, another question examines the role of the specific QE analytic method (propensity score matching, weighting, etc.) in minimizing QE bias (Bifulco, 2012; Cook, Steiner, & Pohl, 2009; Fortson et al., 2012).

The present paper reports findings from a within-study comparison study by combining student-level data from an ongoing longitudinal RCT that evaluates early college high schools in North Carolina (Edmunds et al., 2017; Edmunds et al., 2020) with rich administrative data from North Carolina that include pre- and posttreatment longitudinal information on students who did not participate in this intervention. Our analyses contribute to all three of the open questions concerning the WSC literature listed above. First, we examine three outcomes with natural pretests (English 1 test scores, high school absences, and ACT scores) and three that we consider lacking natural or matching pretests (9th-grade retention, being on-track for college in twelfth grade, and high school graduation). Second, we inform the ongoing discussion regarding the extent to which imposing locational restrictions on the composition of QE comparison groups reduced/eliminated QE bias by implementing two sets of QE approaches. One set restricted QE comparison students to come from the same feeder middle schools as the treatment students ("local" analyses) while the other set did not impose any such restrictions ("global" or "statewide" analyses). Finally, we implemented various propensity scoring techniques (nearest neighbor matching, radius matching, and weighting) to compare the roles of these analytic techniques in reducing QE bias.

We found that for the three outcomes we consider to have natural pretests, multiple QE models replicated empirical benchmarks. For the three outcomes we considered to lack natural pretests, the results were less encouraging. For high school graduation, only one model yielded a sufficiently close QE estimate to the benchmark. For retained in ninth grade, none of the QE models replicated the experimental estimate, and for being on track for college, the imprecision of the QE estimates led to indeterminacy.

In addition, we found that the statewide QE models had smaller (in absolute value) biases than local models for all six outcomes. It was also striking that statewide models replicated the experimental benchmarks for two outcomes (absences and ACT scores) for which local models performed very poorly. An important feature of the early colleges is that they are schools of choice and most attract many more applicants than they can enroll. This suggests that student-level factors may drive the selection process for this intervention and local QE models, which substantially limited the pool of potential comparison students, may have distorted balance on unobservable confounders while prioritizing balance on unobservable covariates. This underlines the principle that QE models should carefully consider the selection processes, local conditions, and the possibility that imposing geographic restrictions may in same cases cause harm more than help.

Among the different QE analytic techniques we implemented, propensity score weighting tended to outperform the other methods in local analyses. For statewide models, a specific method did not stand out in terms of yielding better correspondence. Finally, the direction of the QE bias was generally positive, i.e., QE estimates tended to favor the early colleges more than the experimental benchmark. This result is consistent with the existence of unobserved confounders that are positively associated with attending an early college and outcomes we examined.

The rest of the paper is structured as follows. The following section provides background information on the ECHS initiative as implemented in North Carolina, the early college RCT used in this WSC, and a summary of WSCs conducted in education. The third section presents an overview of data sources and measures. The fourth section describes the design of the WSC and introduces our statistical

framework. The fifth section presents the results from the WSC and the sixth section provides a discussion of the implications of our findings.

BACKGROUND

This section of the paper provides background information on the early college initiative in North Carolina. It also summarizes the existing education WSCs that have assessed the commonly used QE design and analysis features (e.g., employing matching to construct comparison groups and using multivariate regression models to estimate program impacts) that are relevant for the QE methods examined in this paper.[1]

Early College High Schools (ECHS)

Early colleges are small schools (that typically enroll between 100 and 400 students) primarily located on campuses of two- or four-year colleges or universities. Students can earn, at no financial cost to them, up to two years of transferable college credit or an associate degree while simultaneously satisfying state high school graduation requirements. Early colleges are designed to ease the transition from high school to college for students who face barriers on the path to enrolling in college (Roderick et al., 2009). As part of their mission, early colleges seek to serve historically disadvantaged populations, including first-generation college students and students at-risk of dropping out of high school.

Like magnet or charter schools, students choose whether to apply to an ECHS, so these schools have no set admission pool although generally only students from the host county may apply. Many early colleges have slots for all who apply, though some are oversubscribed. In these cases, lotteries are often, but not always, used to select which of the applicants will be invited to enroll. Many schools conduct screening interviews with students and their families. Due to the rigorous nature of the curriculum at early colleges, 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 economically and academically disadvantaged youth *and* students prepared to succeed in college-level coursework—combined with the fact that these are schools of choice, raises the strong possibility of differences in the student populations served by early colleges and traditional public high schools. To the extent that priorities and recruitment techniques differ across sites, it is also possible that these student background differences themselves could even differ across ECHS sites.

Nationally, there are over 240 early colleges in 28 states. North Carolina (NC), with its strong community college and state university systems, is home to 78, which is approximately 30 percent of all ECHSs in the nation, and more than any other state. Each ECHS in North Carolina currently receives a $310,000 grant in addition to the standard per-pupil funding from state, local, and federal sources. In total, the North Carolina General Assembly allocates more than $20 million in additional funding to support this innovative educational approach. Figure A1 in the Appendix at the end of this article shows that about two-thirds of North Carolina counties have an ECHS and that they are spread across all regions of the state. Early colleges that are part of an existing lottery study (Edmunds et al., 2010, 2012) and those that are not part of an existing lottery study are in all parts of the state, but there are very few lottery study participant early colleges in the coastal plain (eastern North Carolina).

[1] The related line of research that uses WSCs to assess the internal validity of regression discontinuity designs (RDDs) (e.g., Chaplin et al., 2018) is not included in this discussion as RDDs are not relevant for this paper.

Early colleges spread rapidly under the auspices of North Carolina New Schools (NCNS), a nonprofit organization that supported early colleges and STEM-oriented high schools in North Carolina, with seed funding from the Gates Foundation. NCNS guided early colleges in North Carolina to implement a core set of design principles: college readiness, powerful teaching and learning, personalization, redefined professionalism, leadership, and purposeful design (Edmunds et al., 2013). A unique feature of NC early colleges is that this intermediary organization delivered initial and ongoing technical assistance to staff starting ECHSs on how to implement these design principles, which increased the fidelity of the intervention relative to what it might be without intensive technical assistance. After NCNS filed for bankruptcy in May of 2016, plans for supporting ECHS sites were picked up by other entities including the North Carolina Department of Public Instruction (NCDPI).

Edmunds et al. (2013) theorize that the success of early colleges stems in part from a school culture of "mandated engagement," which permeates relationships among students, teachers, and administrators. As new small schools of choice for both students and teachers, designed around a shared mission, early college staff include highly committed teachers who believe in the mission and design principles and students who were recruited in part based on the mission. Early colleges raise academic rigor by enrolling students in college-level courses starting in freshman year. To help students meet these higher expectations, early colleges are staffed with teachers, counselors, and administrators who understand that personalization and academic support are critical for student and organizational success.

Case studies and survey research provide a flavor of the unique organizational culture of early colleges. They highlight caring relationships, support, academic identity, and high expectations (McDonald & Farrell, 2012). Students report that they felt prepared for postsecondary education, valued relationships with teachers, and benefited from the small learning communities (Edmunds et al., 2010, 2012; McDonald & Farrell, 2012). Survey analysis reveals that relative to students in traditional public high schools, ECHS students reported statistically significantly higher levels of expectations, more rigorous and relevant instruction, better staff-student relationships, and more frequent and varied types of support. Effect sizes ranged from 0.37 to 1.07, computed on mean differences in survey responses between students who entered an ECHS lottery and were randomly assigned to treatment and control groups (Edmunds et al., 2013).

The Early College RCT in North Carolina

The RCT that provided the empirical benchmarks used in this paper is an ongoing prospective study covering 19 early colleges in North Carolina.[2] The study sample includes more than 4,000 students who applied to one of the 19 participating schools in eighth grade between the 2005/2006 and 2010/2011 school years. The research team implemented lotteries to divide the applicants into two groups: those offered admission (treatment group) and those denied admission (control group). The majority of 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). The study has reported evidence of treatment-control equivalence on baseline characteristics, suggesting successful randomization (Edmunds et al., 2012).

[2] Details on the study design and results can be found in Edmunds et al. (2012, 2013, 2017, and 2020).

Lessons Learned from Existing Within-Study Comparisons in Education

Although earlier WSC studies in education reported weak correspondence in RCT- and QE-based impact estimates (Agodini & Dynarski, 2004; Wilde & Hollister, 2007), more recent WSCs that had access to larger and more diverse sets of potential comparison group members and more extensive sets of potential covariates reported very similar experimental and QE impacts on test score outcomes (e.g., Abdulkadiroglu et al., 2011; Bifulco, 2012; Cook et al., 2020; Dong & Lipsey, 2018; Fortson et al., 2012; Steiner et al., 2010). Following the highly influential qualitative synthesis of the existing WSCs at that time by Cook, Shadish, and Wong (2008), most of these recent WSCs typically go beyond the question of whether it is possible to replicate experimental impact estimates via QE methods and examine the role of three design and analysis features inherent to QE approaches in the bias of the resulting estimates: (1) whether any locational or setting-based restrictions were imposed for the selection of comparison group members, i.e., whether the comparison cases were *local* as they were drawn from the same locations or settings as the treatment cases (Bifulco, 2012; Cook et al., 2020; Wong et al., 2017); (2) whether the covariates used to account for the nonrandom selection of cases into treatment were *focal*, i.e., good predictors of selection into treatment and outcomes of interest (Cook et al., 2020; Dong & Lipsey, 2018; Hallberg et al., 2018); and (3) the specific statistical or econometric analysis techniques used in the construction of the comparison group or modeling the relationship between the outcome and program participation to estimate the program effects (Bifulco, 2012; Fortson et al., 2012). Below we describe these studies in detail and outline the existing gaps in the line of research that have motivated our study.

Bifulco (2012) measured the bias produced by a dozen quasi-experimental approaches using an experimental study of the impact of attendance in a magnet school on children's reading performance. This work suggests that the pool from which comparison units are drawn for the QE analyses has a substantial impact on the accuracy of the replication. In this study, drawing comparison cases from the same districts or districts with similar student characteristics substantially reduced bias. When comparisons were drawn from districts with different student characteristics than the treatment students' districts, the addition of pretreatment test scores to a set of existing demographic covariates that include race/ethnicity and socioeconomic status was insufficient to reduce treatment selection bias.

Three evaluations of charter schools conducted by researchers from Mathematica Policy Research have included sub-studies to validate QE models. All of these studies reported close correspondence between RCT and QE estimates and included very similar QE models: focal covariates, including baseline test scores, and local matching from feeder elementary and middle schools (Fortson et al., 2012; Furgeson et al., 2012; Tuttle et al., 2013). One of these studies (Fortson et al., 2012) tested the validity of four QE methods: (1) OLS regression modeling that controlled for pretest measures and demographic characteristics; (2) exact matching on a specified set of baseline characteristics including grade level, demographics, and pretest; (3) propensity score matching using the pretest and demographic characteristics and higher-order terms and interactions between the baseline characteristics; and (4) fixed effects modeling. This study found that OLS regression modeling yielded estimates that were statistically significantly different from the experimental benchmarks and they led to a different policy conclusion (positive program effects) than the benchmarks (null effects). The other QE approaches, however, produced QE estimates that were not statistically distinguishable from the RCT benchmarks.

Hallberg et al. (2018) assessed the role of the pretest measure of the outcome in the reduction of bias inherent to observational studies summarizing results from three within-study comparisons. Their analysis suggests that controlling for one pretest

measure would substantially reduce QE bias, using two waves of a pretest is expected to reduce bias more than a single pretest, and employing a large and heterogeneous set of covariates that includes one or more pretest measures is likely to perform the best.

A recent WSC (Cook et al., 2020) tested the role of the three QE design elements (local matching, using a pretest measure of the outcome as a covariate, and using a rich set of multidimensional covariates other than the pretest) in reducing QE bias in the evaluation of a prekindergarten mathematics curriculum. This study found that the QE model that combined all three elements yielded the minimum bias (less than .10 standard deviations) and nearly all bias reduction was due to local matching and not to the pretest or other covariates.

Finally, Wong, Valentine, and Miller-Bains (2017) conducted a qualitative synthesis of 12 within-study comparisons in education that used achievement outcomes. They summarized the empirical evidence on the role of three types of covariates and statistical controls—pretest measures, local geographic matching, and rich covariates with a strong theory of selection—in bias reduction in QE studies. They conclude that the pretest can substantially reduce bias and almost completely eliminate it when it is highly correlated with the outcome and selection into treatment and has a linear baseline trend for both groups (no adjustment for trend effects is needed in that case). Some bias remains in cases where there are differential baseline trends for the treatment and comparison units (i.e., selection is based on differences in baseline trends) or there are other important selection covariates. In those cases, trend effects or additional selection covariates should be controlled for to reduce bias. They did not find an added advantage of local comparison group matches over nonlocal matches when the treatment and comparison groups are balanced on covariates. However, they found that using local comparison cases that differ from the treatment cases on covariates may lead to substantial bias. They also noted that observational methods perform well when used with a rich covariate set organized around a unifying theory of factors that may be related to selection into treatment.

This literature demonstrates that QE impacts on test score outcomes from authentic educational settings can have high internal validity. However, there are still some open questions: To what extent does imposing geographic restrictions on the QE comparison groups matter? When used with a comprehensive set of focal covariates and large number of potential comparison students, do QE methods that differ by how selection into treatment is modeled and how many students included in the comparison group yield different answers? Building on the existing WSCs in education, the present study tackles these questions by assessing the bias of a variety of QE methods that differ by the locational restrictions placed on potential comparison group members, the propensity score model specifications, and how propensity scores are used to construct the comparison groups. As noted earlier, two important contributions of this paper to the research base in education WSCs are assessing the performance of QE methods when outcome measures and covariates exclusively come from administrative SLDSs and for one-time outcome measures that do not have natural pretests.

DATA SOURCES AND MEASURES

We use a rich longitudinal student-level data set constructed from administrative elementary and secondary public school data from the North Carolina Department of Public Instruction (NCDPI). These data include the full population of students who attended any public school in North Carolina during the 2004/2005 to 2015/2016 school years, and individual students only become unobserved if they

leave the public system. For this paper, we focus on six high school outcomes—English I test scores, average attendance through high school, ACT test scores (administered to all eleventh graders in North Carolina since 2012), 9th-grade retention, being on track (or prepared) for college in twelfth grade,[3] and five-year high school graduation. This is a comprehensive set of outcomes that not only represents important academic and engagement measures for high school students but also includes potential predictors of longer-term outcomes such as attainment of postsecondary credentials, employment, and wages.

In addition to data on student outcomes, the data set includes many student and school-level variables measured prior to entry into high school that can be used as covariates to control for potential confounding. At the student level, the data include demographic variables, such as student ethnicity, gender, economic disadvantage, old for grade (defined based on students' age and current grade level), limited English proficiency, disability status, and gifted identification as well as data on prior performance in middle school including 6th- to 8th-grade math and reading test scores, 8th-grade science test scores, taking and passing Algebra I in middle school, middle school attendance, and mobility during middle school.

Based on the available baseline covariates, we can reasonably argue that three outcome measures—English 1 test scores, average high school attendance, and ACT scores—have natural (i.e., matching) pretests (middle school attendance and test scores). Two outcomes—high school graduation and being on track for college—are one-time events. While we expect that the demographic covariates and middle school achievement measures should be correlated with these two outcomes, we do not consider them to have natural pretests as both measures reflect students' entire high school experiences and may be influenced by potential unobserved traits such as motivation. For 9th-grade retention, we do have a covariate (old for grade) that may be considered as a natural pretest since it reflects retention in elementary and middle school. But it may also reflect other factors such as kindergarten redshirting,[4] therefore we treat 9th-grade retention as an outcome without a natural pretest.

The current data set consists of four cohorts of high school students who entered ninth grade for the first time in the 2007/2008, 2008/2009, 2009/2010, and 2010/2011 school years.[5] These students are expected to graduate from high school between the 2010/2011 and 2014/2015 school years. Table 1 provides an overview of these cohorts. In order to be included in these cohorts, students must have been enrolled in North Carolina public schools in ninth grade and also have been enrolled in eighth grade in North Carolina public schools in the prior school year. This sample restriction is necessary in order to ensure that students have pretreatment demographic and performance data. 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 of those who were not enrolled in North Carolina public schools during one of the two years and students who were

[3] The on-track or college readiness outcome is defined as taking and succeeding in the courses that students would need for college. See Edmunds et al. (2017) for a detailed description of how this measure is constructed.

[4] Kindergarten or academic redshirting is the practice of delaying age-eligible kids' enrollment to kindergarten. Its primary aim is to allow for further social-emotional, academic, and physical growth (Katz, 2000).

[5] The 2005/2006 and 2006/2007 cohorts of the RCT (which included about 400 students who applied to two early colleges) are excluded from the WSC due to issues with obtaining baseline data for the potential comparisons in these years. The WSC analysis includes 3,473 students from the 2007/2008 through 2010/2011 cohorts included in the RCT.

Table 1. Cohorts included in study.

	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014
8th grade	1	2	3	4				
9th grade		1	2	3	4			
10th grade			1	2	3	4		
11th grade				1	2	3	4	
12th grade					1	2	3	4
Postsecondary/13th grade						1	2	3
Postsecondary/14th grade							1	2
Postsecondary/15th grade								1

Notes: This table shows the grade progression of the four student cohorts included in the study sample. Each cell corresponds to a school year and grade level combination and the number in a given cell shows the student cohort covered by cell. For example, the first student cohort consisted of students who were in eighth grade during the 2006/2007 school year.

retained in either eighth or ninth grade as well as a small number of students who could not be matched across time based on name, birthdate, and other administrative identifying variables.

In addition to the exclusion of students who do not appear in both eighth and ninth grades, some students are not included in some analyses due to either attrition from the sample over time or missing data. Because the study utilizes administrative data, attrition rates are fairly low. Students who leave the sample may have dropped out of high school, transferred out of the North Carolina public schools, or failed to be matched in subsequent years. Attrition in each year of high school is 7 percent or less and appears similar across the ECHS and non-ECHS students.[6]

Missing data rates are also fairly small. Overall, relatively few students are missing outcome data for the outcomes included in this paper. Missingness for the outcomes ranges from 0 percent missing for five-year graduation rate[7] to 9 percent of students missing English I test 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. Missing covariates are imputed for these students using the "dummy variable" method that entailed (1) replacing missing values for a given covariate with the sample mean and (2) including an indicator for the imputed records in the propensity score and impact estimation models (Stuart, 2010).

The full data set includes a total of more than 450,000 students across the four cohorts across the entire state, but the size of the analytic samples varies for different analyses due to analysis details such as matching techniques (which are described in more detail in the next section). The WSC includes 19 ECHSs, with an original randomized sample that included a total of 3,473 students of whom 2,044 were randomized into treatment and 1,429 were randomized into control. Overall, the compliance rate is 90 percent, with 1 percent of crossovers (students who were assigned to the control group but ended up enrolling in the ECHS to which they applied) and 9 percent no-shows (students who were assigned to the treatment group but did not enroll in any ECHS). The comparison group in the within-study comparison ranges from about 2,266 to 409,185 students, depending on the QE analysis method. Characteristics of these samples are described in detail below when we discuss results.

DESIGN OF THE WSC AND STATISTICAL METHODS

A naïve comparison of the outcomes of students who applied and were accepted to an ECHS to students who did not apply to an ECHS and enrolled in a regular high school would provide a misleading picture of the effect of attending an ECHS due to baseline differences between the two groups. First, these two types of students may have different motivational and cognitive characteristics as well as parental involvement and support, which may be directly related to differences in their interest in ECHS and their high school and postsecondary outcomes. We will refer to such confounders as "individual self-selection factors." Second, the two student types may have been raised in different neighborhoods and had different elementary and middle school experiences. For instance, early college applicants may have had more

[6] Between ninth and twelfth grades, we lost about 9 percent of the treatment sample as opposed to 19 percent of the potential comparison students. The weighted attrition rate for the comparison group, which is calculated using the propensity score weights and constitutes a better counterpart for the treatment group, is 11 percent.

[7] We only have lists of high school graduates in each year. A student who is included in our sample but did not appear in these lists was coded as zero for not graduating.

academic support during middle school that helped them prepare for high school and pursue postsecondary education thereafter. We will refer to these confounders as "geographical or locational factors." Both sets of confounders may also account for differences in the outcomes of interest between the two groups, which should not be attributed to attending an ECHS.

Utilizing data from the RCT, the WSC explores whether it is possible to replicate the experimental impact estimates that are not subject to any biases due to the confounders described above using QE estimators. These replication exercises replace the control group of the RCT with a selected group of non-ECHS students such that the treatment group of the RCT and the resulting comparison group are balanced to the extent possible in terms of the two types of confounders described above. This section of the paper describes in detail the design of the WSC and statistical properties of the QE estimators.

Assessment of WSC Assumptions

Wong and Steiner (2018) present a comprehensive theoretical framework for the design and implementation of different types of WSCs. Per their definition, we conduct a "dependent simultaneous WSC," in which the early college RCT constitutes the benchmark and the observational comparisons are obtained from the administrative databases in North Carolina. This design is considered dependent as the RCT and QE analyses share the same treatment group. For dependent simultaneous WSCs, Wong and Steiner (2018) list the following four assumptions to ensure that the RCT and QE analyses identify the same treatment effect:

1. There is no interference between units in the RCT and QE analyses, i.e., the RCT and QE members' potential outcomes depend only on their treatment assignment status but not on the assignment of others.
2. There are no systematic differences between the QE and RCT control conditions.
3. Potential outcomes are independent of treatment assignment in the RCT, i.e., the RCT produces internally valid effect estimates.
4. In the QE analysis, participants' potential outcomes are independent of the treatment assignment status conditional on observed covariates.

We argue that the first three assumptions are highly plausible in our WSC because the QE comparison group students did not participate in the RCT, were not exposed to the treatment, and had similar high school experiences as the RCT control group. In addition, identical outcome measures were collected from the QE comparison students using identical procedures to those used in the RCT. Therefore, if a given QE estimator does not replicate the experimental benchmarks in our WSC, we will conclude that the fourth assumption must have been violated because of unobserved confounders in the QE analyses.

Estimation of the Experimental Benchmarks

Two features of the longitudinal RCT that provided the experimental benchmarks for the WSC are important for the estimation of these benchmarks. First, the presence of no-shows and crossovers led to at least two different estimands that could be estimated with experimental data: intent-to-treat (ITT), which represents the impact of receiving the offer to enroll in an ECHS and treatment-on-the-treated (TOT) or local average treatment effect (LATE), which captures the effect of actually enrolling in early colleges on students who complied with the random assignment results. Wong and Steiner (2018) argue that for experimental studies with non-compliance, ITT

is the causal estimand of interest because compliers cannot be distinguished from always-takers in the treatment group and from never-takers in the control group, which complicates the identification of the LATE/TOT in the QE arm of WSC. Following their advice, the current WSC analyses use the ITT estimates as the empirical benchmarks.

The second issue stems from the fact that some schools identified priority populations (e.g., first-generation college attendees) for their incoming cohorts. To include these schools in the analysis, the research team stratified the eligible pool of applicants by the priority characteristics and more students from the priority groups were assigned to the ECHS through these stratified lotteries, which led to unequal probabilities of treatment assignment within the study sample. This should be taken into account when calculating experimental benchmarks; otherwise, the resulting treatment and control groups in the RCT would not be balanced in terms of the characteristics used for stratification. For example, stratifying the applicants by their first-generation status and assigning a larger proportion of first-generation students to the treatment group would lead to a higher proportion of first-generation students in the treatment group than the control group.

For the purposes of the WSC, we adopted a weighting strategy that took into account the stratified lotteries. When obtaining the experimental benchmarks, this strategy involved weighting each treatment student by 1 and weighting each control student by the ratio of his or her probability of getting into the treatment group to the probability of getting into the control group. Using these weights allowed us to balance the treatment and control groups on the stratification characteristics by overweighting the control students in strata where higher proportions of the students were assigned to the treatment group and underweighting the control students in strata where lower proportions were assigned to the control group. An advantage of this weighting is that it is directly relevant to the QE arm of the WSC where all treatment students were weighted by 1 and the control students were weighted according to their matching frequency. This ensures that the RCT and QE estimands are comparable.

These weights are used in the following model to estimate the experimental benchmarks:

$$Y_i = \gamma_0 + \gamma_1 T_i + \sum_{l=1}^{L-1} \gamma_{(1+l)} S_i^l + \sum_{m=1}^{M} \gamma_{(L+m)} X_i^m + \varepsilon_i, \quad (1)$$

where:

Y_i = outcome measure for student i.

T_i = treatment indicator for student i, which equals one if student i is randomized to the ECHS group and zero otherwise.

S_i^l = indicator variable for the lottery l, which equals one for students who participated in lottery l and zero for other students ($l = 1 \ldots L$).[8]

X_i^m = m-th covariate for student i. Note that the model controlled for all of the covariates used in the QE analyses described below.

ε_i = random error term for student i.

The coefficient γ_1 on the treatment indicator denotes the experimental impact estimate. We clustered the standard errors at the high-school level (early colleges for treatment students and regular high schools for control students) to account for the potential clustering of student outcomes within schools.

[8] Per our definition, a lottery includes all students who applied to enroll in an early college in a given year. As explained before, some lotteries were stratified based on priority characteristics determined by early colleges.

Properties of Quasi-Experimental Estimators

We compared the RCT-based impact estimate for each outcome with estimates from a variety of QE models. The QE estimators differed by how the comparison groups were constructed but they all used the same set of student-level covariates to account for the potentially systematic differences between students who enrolled in early colleges through lotteries and students who did not participate in those lotteries. These covariates, all of which were measured before the treatment commenced, include:

- *Demographics*: Gender, race/ethnicity, socioeconomic status, limited English proficiency status, disability status, whether the student is gifted, mobility in middle school (as a proxy for family stability), whether the student is old for grade;[9]
- *Middle school academic achievement*: Averages of 6th- through 8th-grade state test scores in reading and math,[10] 8th-grade test scores in science, and taking and passing Algebra 1 in eighth grade; and
- *Attendance in middle school*: Average absenteeism in sixth through eighth grade is used as a proxy for motivation and academic engagement.

Students who attend early colleges tend to be high performing and highly motivated students who have postsecondary aspirations in middle school. The combination of the covariates listed above would be expected to capture most of these traits, especially the individual self-selection factors described above. Notable omitted variables that could potentially act as confounders through their joint relationship with selection into early colleges and outcomes we examined include parental engagement and students' and their parents' perceptions about the value of postsecondary education. Such characteristics, however, are rarely available in administrative data sources that are typically accessible to education researchers.

The covariates we used cover the typical covariates available to education researchers in extant databases; therefore, this paper provides a fair assessment of the possibility of replicating experimental estimates with such secondary education data found in administrative data sets.

The QE estimators we assessed varied across the following dimensions:

- Geographic restrictions placed on potential comparison group members (local vs. statewide);
- Whether and how propensity scores were utilized in the analysis (OLS regressions that do not use propensity scores, propensity score weighting, or propensity score matching).

Table 2 summarizes the features of the different QE estimators we used, which are described in more detail in the following subsections.

Identifying Potential Comparison Group Members

Existing WSCs are inconsistent about the role of imposing locational restrictions on the selection of potential comparison group members. Some WSCs showed that choosing comparison units from "local" untreated units that share the same school,

[9] A student could be old for grade if he or she was retained in a prior grade or because of kindergarten redshirting.

[10] We considered controlling for 6th-, 7th-, and 8th-grade math and reading test scores separately to capture individual achievement trajectories through middle school but this led to higher missing rates for these measures.

Table 2. Quasi-experimental models.

Label	Geographical Restriction for Potential Comparisons	Propensity Score Estimation	Details on Matching	Additional Controls In Impact Regressions
Local OLS	Non-ECHS students (excluding those in the RCT control group) from same feeder middle schools as treatment students	N.A	N.A.	Cohort by feeder middle school interactions
Local 1-to-1 Matching		Probit	1-to-1* with replacement	
Local 4-to-1 Matching			4-to-1* with replacement	
Local Radius Matching			Radius*	
Local Propensity Weighting			N.A.	
Statewide OLS	All non-ECHS students (excluding those in the RCT control group) in NC	N.A	N.A.	Cohort indicators
Statewide 1-to-1 Matching		Probit	1-to-1* with replacement	
Statewide 4-to-1 Matching			4-to-1* with replacement	
Statewide Radius Matching			Radius*	
Statewide Propensity Weighting			N.A.	

Notes: *Local 1-to-1, 4-to-1, and radius matching estimators implemented exact matching on cohort and feeder middle schools while statewide matching estimators implemented exact matching on only cohort.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

neighborhood, or district with the treatment units was critical for replicating experimental results (e.g., Bifulco, 2012; Cook, Shadish, & Wong, 2008; Steiner et al., 2010). Three WSCs included in the Wong, Valentine, Miller-Bain (2017) synthesis, however, suggest that such locational restrictions were not influential for bias reduction when the treatment and comparison groups are balanced on focal covariates. Furthermore, they showed that such restrictions may limit the pool of potential comparison groups and yield inadequately balanced treatment and comparison groups, thereby leading to more biased estimates than QE approaches without such restrictions that can form more tightly balanced groups. Therefore, the bias implications of imposing any geographical or locational restrictions on the construction of the QE comparison group is still an open question.

In the case of early colleges, students who attended middle schools that emphasized postsecondary education may have been more likely to pursue postsecondary education and more likely to apply to an ECHS than their peers without such supports. However, these locational factors may not be fully captured by the student-level covariates available in the extant data. To examine the extent to which controlling for such factors was instrumental for reducing bias, we used a set of QE approaches that implemented a variant of local propensity score analysis as follows: For the QE models that used propensity score matching, we conducted the matching process separately within blocks where each block included treatment and potential comparison students from the same cohort who attended the same middle school (*local matching*). For those that used propensity score weighting, non-ECHS students who attended different middle schools than treatment students were dropped from the analyses (*local weighting*).

On one hand, this restriction may allow us to account for locational confounders. On the other hand, it may yield groups that are unbalanced on unobserved/omitted student-level variables because it forces us to compare students who applied to an ECHS with those who did not apply despite the two groups attending the same middle school, which presumably led them to have similar exposure to institutional factors that would influence students' and their parents' motivation towards pursuing postsecondary education. To examine the role of feeder middle schools in the self-selection of students into early colleges and the possibility of imposing locational restrictions yielding treatment and comparison groups that are *balanced on observables but inadequately balanced on omitted variables*, we tested an additional set of QE models that implemented *global or statewide* propensity score analyses such that no restrictions with respect to the feeder middle schools were imposed on the QE comparison groups. That is, potential comparison groups for the statewide analyses included all non-ECHS students from the relevant 9th-grade cohorts in North Carolina. Contrasting results from these analyses with results from local models allows us to assess the benefits and potential drawbacks of imposing locational restrictions on QE comparison groups.

All QE comparison groups excluded non-ECHS students who were in the original control group of the RCT so there is no overlap between the QE comparison groups and the experimental control group.

Estimation of Propensity Scores

Propensity scores were estimated using probit models specified with the covariates listed above. Separate models were estimated for local and statewide analyses. Propensity score estimates capture the probability of applying and receiving the offer to attend an ECHS conditional on the covariates included in the model.

Details on How Estimated Propensity Scores Were Used: Matching and Weighting

Matching is the most common application of propensity score analysis, with many variants (Stuart, 2010). For parsimony, we are reporting results from three matching methods:[11]

1. *One-to-One matching*: each treatment student is matched with one potential comparison student with the closest propensity score within the pre-specified caliper (±0.2 of the standard deviation [SD] of the propensity score set per Stuart, 2010).
2. *Four-to-One nearest neighbor matching*: each treatment student is matched with the closest four comparison students within his or her caliper (±0.2 of the SD of the propensity score).
3. *Radius matching*: each treatment student is matched with all potential comparison students whose propensity scores are within the specified caliper of his or her score (±0.2 of the SD of the propensity score).

In all cases, a comparison student can be matched with multiple treatment students (matching with replacement) and the frequency of being used as a matched comparison was captured via weights. Treatment students who did not have any comparison students within their caliper were unmatched and excluded from the estimation of early college effects. This is a version of enforcing "common support," which is used by some propensity score applications to ensure the overlap of the range of the propensity scores between the treatment and matched comparison groups (Caliendo & Kopeinig, 2008; Garrido et al., 2014). We track the number of unmatched treatment students because the exclusion of a large proportion of treatment students from the QE analyses can raise concerns about the comparability of the experimental and QE estimands.

These methods allow us to assess the potential bias-precision trade-off between balance of the treatment and comparison groups and effective sample size. One-to-one matching takes the "best match" for each treatment student within the specified caliper; therefore, it places a higher priority on generating closely matched treatment and comparison pairs to minimize bias. Radius matching, on the other hand, uses all potential comparison students within the specified caliper, placing a higher priority on maximizing the size of the comparison group and precision of the effect estimates, but this can come at the cost of less balanced groups and more bias. Four-to-one nearest neighbor matching is a more balanced approach as it does not prioritize bias or precision as strongly as the other approaches.

As an alternative to matching, we used the estimated propensity scores to create weights (propensity weighting or PW). Following Stuart (2010), treatment students were weighted by 1, comparison students were weighted by $\frac{\hat{P}}{1-\hat{P}}$ (i.e., odds of selection) where $\hat{P}$ is the estimated propensity score. An advantage of weighting over the three matching approaches is that the analysis retains all treatment and potential comparison students.

[11] Stuart (2010) suggests 0.2 standard deviations (SD) as a reasonable caliper for propensity score analyses but beyond that, there is little empirical or theoretical guidance for choosing an optimal caliper. We also implemented 1-to-1 matching without replacement, 1-to-1 matching without a caliper, 4-to-1 matching without a caliper, and all three methods with a narrower caliper (±0.1 of the standard deviation of the propensity scores). These methods yielded similar results to those discussed in the paper. These additional results are available upon request.

Assessing Quality of Matches

Following Rosenbaum and Rubin (1985) and What Works Clearinghouse (2018), we assessed the quality of the matches using standardized treatment-comparison differences (aka effect sizes) calculated as follows. For each covariate, we first fit a weighted regression model that used the covariate as the dependent variable, and the treatment group indicator and indicators for cohort by feeder middle school interactions for local models and cohort indicators for statewide models as independent variables. The standardized difference was then calculated as the ratio of the coefficient on the treatment indicator to the pooled standard deviation of the covariate across the treatment and potential comparison students. We required the standardized differences to be less than 10 percent of a SD in absolute value for all covariates. Our threshold is more stringent than the 0.25 SD threshold used by the WWC.

Estimation of the QE Effects

The following model was used to estimate the ECHS effect:

$$Y_i = \pi_0 + \pi_1 T_i + \sum_{b=1}^{B-1} \pi_{(1+b)} I_i^b + \sum_{m=1}^{M} \pi_{(B+m)} X_i^m + \varepsilon_i, \quad (2)$$

where:

Y_i = outcome measure for student i.

T_i = treatment indicator for student i, and equals one if student i is an ECHS student and zero otherwise.

I_i^b = indicator variable for the b-th analysis block for student i. It equals one if student i is a member of the b-th block and zero otherwise. As shown in Table 2, local analyses used interactions between cohort indicators and feeder middle schools as analysis blocks while statewide analyses used cohort indicators as analysis blocks.

X_i^m = m-th covariate for student i. We controlled for all of the covariates used in the estimation of the corresponding propensity score to increase the precision of the QE impact estimates and be doubly-robust (Bang & Robins, 2005).[12]

ε_i = random error term for student i.

The coefficient π_1 denotes the estimated ECHS effect. Standard errors were clustered at the high-school level (early colleges for treatment students and regular high schools for comparison students). In addition to the QE models that utilized the various propensity score analysis methods described above, we also estimated ECHS effects using "naïve" OLS models that used all potential comparison group members for local and statewide analyses. These models were specified as in equation (2) but essentially weighted all potential comparison group students by 1. These analyses yielded 12 QE impact estimates for each outcome measure.

Assessing Which QE Models Replicated Experimental Results

The final step of the WSC study was to assess which (if any) of the 12 QE effect estimates replicated the RCT-based benchmark for each outcome. Historically, the

[12] Using the baseline characteristics in the matching process and using them as covariates in the estimation of impacts gives the analyst two chances to get the "right" model specification (once in the propensity model and another time in the impact model for the outcome measure). Therefore, these estimators are called "doubly-robust."

WSC studies used different approaches to do this assessment. For example, Fortson et al. (2012) examined whether the experimental and QE estimates had the same sign and statistical significance, and similar magnitudes (i.e., both estimates led to the same policy conclusions). Hill, Reiter, and Zanutto (2004) required the 95 percent confidence intervals of the two estimates to overlap while Hallberg et al. (2018) required the difference between two estimates (i.e., bias in the QE estimates) to be less than 0.15 standard deviations and not statistically significant (assessed using the bootstrapped standard error of the difference).

More recently, Steiner and Wong (2018) proposed a comprehensive framework to assess the correspondence between experimental benchmarks and QE estimates. This framework entails formally assessing the insignificance of the difference between two estimates ("insignificant difference") and statistical equivalence of the two estimates ("significance of equivalence"). The null hypothesis for the first assessment states that the QE bias is zero, i.e., $H_0^d: \ bias_{QE} = \pi_1 - \gamma_1 = 0$ where π_1 is the QE effect from equation (2) and γ_1 is the RCT effect from equation (1). Failing to reject the null hypothesis indicates that the difference between the QE and RCT effect estimates is statistically insignificant, i.e., providing support that the two effects are equivalent.

It is important to note that one may fail to reject the null hypothesis above if the precision of the QE or RCT effect estimates are low even when $bias_{QE}$ is sizable.[13] Therefore, we complement this assessment with an additional assessment that formally tests the equivalence of the two effects. This test uses a composite null hypothesis that states that the difference between the two estimates is larger than a threshold, δ_E, which accounts for the fact that the point estimates of the two effects could slightly differ because of sampling error: $H_0^e: \ |\pi_1 - \gamma_1| \geq \delta_E$. Rejecting this null hypothesis suggests that the difference between two effects is negligible, which provides statistical support for the equivalence of the effects.

Steiner and Wong (2018) conceptualize the composite null hypothesis H_0^e as two one-sided hypotheses: $H_{01}^e: \ \pi_1 - \gamma_1 \geq \delta_E$ and $H_{02}^e: \ \pi_1 - \gamma_1 \leq -\delta_E$. Rejecting both of these null hypotheses suggests that the two effects are equivalent. Failing to reject at least one provides evidence that the two effects are not equivalent.

The statistical correspondence of the QE and RCT effects is determined by these two assessments, as Figure 1 illustrates in the four possible scenarios. "Equivalence" is indicated if both assessments suggest correspondence (i.e., H_0^e is rejected but H_0^d is not) and "Difference" is indicated if both assessments point to noncorrespondence (i.e., H_0^d is rejected but H_0^e is not). If the equivalence test supports correspondence but the difference test does not, this is considered to be "Trivial difference." This may happen if both the QE and RCT estimates are highly precise or δ_E is large so even a small difference between QE and RCT effects is detected. Finally, "Indeterminacy" captures cases where the difference test supports correspondence, but the equivalence test does not. This may occur when either test does not have sufficient power because of small sample sizes and imprecise effect estimates.

For both assessments, we set the significance level (α) to 0.05 and we used bootstrapping (with 500 bootstrapped samples[14]) to account for the covariance between the RCT and QE estimates because the treatment group was used in both estimation procedures. When testing the equivalence of the two estimates, we set δ_E to 0.10 SDs.

[13] This was the case for some of the existing WSC studies in education (Wong, Valentine, & Miller-Bain, 2017).

[14] The bootstrapping procedure accounted for our nested data structure by using nonparametric bootstrapping at the high-school level (i.e., random sampling of high schools *with* replacement and random sampling of students within high schools *without* replacement), which is shown to be optimal with hierarchical data (Ren et al., 2010).

Test of Insignificant Difference between RCT and QE Estimates	**Test of Equivalence of RCT and QE Estimates**	
	Insignificant Equivalence (Noncorrespondence)	Significant Equivalence (Correspondence)
Significant Difference (Noncorrespondence)	**Difference**	**Trivial Difference**
Insignificant Difference (Correspondence)	**Indeterminacy**	**Equivalence**

Notes: This figure is adapted from Table 1 in Steiner and Wong (2018). It shows the four potential conclusions of the correspondence assessment. The rows show the results of the test that assesses whether the difference between the RCT and QE estimates (i.e., QE bias) is statistically significant. A significant difference provides evidence for noncorrespondence while an insignificant difference provides evidence for correspondence. The columns show the results from the test that assesses the equivalence of the RCT and QE estimates by testing whether the difference between the RCT and QE estimates is larger than a threshold that represents a negligible difference (e.g., a tolerable effect size difference that can be caused by sampling error). A significant equivalence result provides evidence for correspondence and an insignificant equivalence result provides evidence for noncorrespondence. Please see the text for a more in-depth description of the underlying hypotheses tested in each assessment.

Figure 1. Correspondence of RCT and QE Estimates.

This threshold[15] (which corresponds to 3 to 4 percentage points for the binary outcomes we examined in this paper) was suggested by Steiner and Wong (2018) and seemed appropriate for our outcomes because education evaluations typically use it as the minimum detectable effect size in power calculations for these outcomes, i.e., 0.10 effect size is considered as substantively meaningful.

RESULTS

We start with describing the characteristics of the samples used in the WSC analyses. Table 3 presents the means of the covariates for three groups of students: treatment students, potential comparison (i.e., non-ECHS) students used in local models, and potential comparison students used in statewide models. Table 3 shows that while treatment and local and statewide comparison groups had similar race/ethnicity and mobility rates, there were considerable differences between the treatment and comparison students on the other characteristics. Treatment students were less likely to have disabilities; more likely to be female, eligible for free/reduced priced lunch, and gifted; and they had higher test scores and attendance rates in middle school. While these differences had similar magnitudes across the two comparison groups for the demographic variables, educational status variables, and attendance rates, statewide comparison students had slightly better average test scores in End of Grade tests in math, reading, and science than the local comparison students.

Table 4 presents coefficient estimates from local and statewide probit regressions that used the treatment indicator (= 1 if treatment student, = 0 if local or statewide potential comparison student) as the dependent variable and the three sets of covariates (demographics, achievement, and attendance) as the independent variables or predictors. Estimates of the probit coefficients shown in Table 4 suggest that in both local and statewide models being female, being eligible for free/reduced price lunch, higher scores on middle school reading, math, and science tests, and passing Algebra 1 in eighth grade were positively associated with being in the treatment group.

[15] It corresponds to about 4 percent for a binary outcome with a mean of 80 percent (such as high school graduation in our sample) and about 3 percent for a binary outcome with a mean of 10 percent (such as 9th-grade retention).

Table 3. Means of covariates for WSC samples.

	Treatment Group (1)	RCT Control Group (2)	Potential Local Comp. Group (3)	Potential Statewide Comp. Group (4)
Demographics				
Male	39.90%	39.63%	51.90%	51.50%
Asian	1.30%	1.41%	1.50%	2.20%
Black	28.50%	26.17%	27.40%	28.80%
Hispanic	8.00%	7.71%	7.60%	9.00%
American Indian	0.30%	0.83%	0.60%	1.50%
Multiracial	3.76%	2.56%	3.20%	3.10%
White	58.20%	61.32%	59.62%	55.36%
8th Grade Free/Red. Price Lunch Eligibility	52.10%	50.86%	49.80%	45.80%
8th Grade ELL Status	3.70%	3.25%	4.50%	5.20%
8th Grade Disability Status	4.20%	5.51%	12.70%	12.90%
8th Grade Gifted Status	21.10%	21.67%	14.80%	16.30%
Old for Grade	11.60%	12.81%	22.50%	20.50%
Moved Middle Schools	24.00%	25.51%	25.90%	25.90%
Cohorts				
1st cohort (8th grade in 2006–07)	12.10%	12.05%	13.20%	25.10%
2nd cohort (8th grade in 2007–08)	23.30%	24.04%	26.30%	25.10%
3rd cohort (8th grade in 2008–09)	35.80%	34.53%	33.00%	25.00%
4th cohort (8th grade in 2009–10)	28.90%	29.38%	27.50%	24.90%
Achievement				
Middle School Avg. Math Score (z-score)	0.26	0.28	-0.12	-0.02
Middle School Avg. Reading Score (z-score)	0.31	0.33	-0.1	-0.02
Passed Algebra 1 in Middle School	22.80%	25.40%	19.60%	22.20%
8th Grade Science Score (z-score)	0.19	0.20	-0.11	-0.07
Absences				
Middle School Avg. Days Absent	6.52	6.82	8.01	7.81
Number of Observations	2,053	1,437	44,073	411,521

Notes: Middle school average test scores and days absent are simple averages of these measures in the sixth, seventh, and eighth grades. A student could be old for grade if he or she was retained in a prior grade or because of kindergarten redshirting.

Being African American, having gifted status, and higher absenteeism are negatively associated with being in the treatment group for the local models while being American Indian, Asian, Hispanic, having a disability, and having lower absenteeism were negatively associated with the treatment in the statewide models.

The coefficient estimates from the probit models were used to create the propensity scores. Table 5 presents an overview of the sizes of the treatment and comparison

Table 4. Coefficients from probit regressions.

Covariates	Local		Statewide	
	Demographics + Achievement + Absences		Demographics + Achievement + Absences	
Number of Observations	46,117		409,185	
	Coeff.	P-value	Coeff.	P-value
American Indian	-0.166	0.32	-0.465	0.00
Asian	-0.119	0.21	-0.261	0.00
Black	0.185	0.00	0.034	0.11
Hispanic	0.088	0.08	-0.080	0.03
Multiracial	0.093	0.11	0.029	0.49
Male	-0.198	0.00	-0.143	0.00
Gifted	-0.108	0.00	-0.021	0.39
Have Disability	-0.085	0.09	-0.158	0.00
8th Grade Free/Red. Price Lunch	0.213	0.00	0.242	0.00
8th Grade ELL Status	0.120	0.09	0.043	0.40
Moved Middle Schools	0.004	0.88	0.010	0.61
Old for Grade	0.002	0.96	-0.010	0.80
Old for Grade * Free Lunch	-0.116	0.08	-0.063	0.21
Cohort 2	-0.099	0.43	0.052	0.58
Cohort 3	-0.017	0.90	0.197	0.04
Cohort 4	-0.006	0.96	0.135	0.15
Middle Sch. Avg. Math Scr. (z-score)	0.179	0.00	0.068	0.00
Middle Sch. Avg. Reading Scr. (z-score)	0.186	0.00	0.110	0.00
8th Grade Science Scr. (z-score)	0.061	0.00	0.076	0.00
Passed Algebra in 8th Grade	-0.303	0.00	-0.239	0.00
Middle Sch. Avg. Absences (days)	-0.008	0.00	0.007	0.00

Notes: Entries in the table show the probit regression coefficients. Middle school average test scores and days absent are simple averages of the same measures in the sixth, seventh, and eighth grades.

groups by analysis method. Weighting analyses utilized all treatment and potential comparison students by the nature of this analytic strategy while matching analyses excluded some treatment and potential comparison students from the estimation of effects because they were not used as matches. Table 5 shows that large proportions of treatment students were included in the matching analyses. Eighty-nine treatment students (4.3 percent) were unmatched in local matching analyses (for lack of any potential comparisons within their caliper) and only nine treatment students were unmatched in statewide matching analyses. Table 5 also shows that percentage of potential comparison students included in the matching methods varied by the analysis approach. Local radius matching utilized half of the potential comparison students while statewide radius matching matched almost all of the potential comparison students with treatment students. As expected, these proportions were much smaller for one-to-one methods at 5 percent for local analyses and 1 percent for statewide analyses.

Next, we assessed the extent to which matching or weighting worked by examining the balance of the matched treatment and comparison groups. We first

Table 5. Overview of matching and weighting.

	Local				Statewide			
	Demographics + Test Scores + Absences				Demographics + Test Scores + Absences			
	1-to-1 Matching	4-to-1 Matching	Radius Matching (0.2 SD)	PW	1-to-1 Matching	4-to-1 Matching	Radius Matching (0.2 SD)	PW
Matched Treatment Students	1,964	1,964	1,964	2,053	2,044	2,044	2,044	2,044
Non-Matched Treatment Students	89	89	89	0	9	9	9	0
% Treatment Students Unmatched	4.3%	4.3%	4.3%	-	0.4%	0.4%	0.4%	-
QE Comparison Group	2,266	6,061	22,216	44,073	3,946	8,205	4,06,733	4,09,185
% Potential Comparisons Matched	5.1%	13.8%	50.4%	100.0%	1.0%	2.0%	99.4%	100.0%

Notes: PW stands for propensity weighting. Radius matching used a caliper of 0.2 standard deviation of the propensity score.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

examined the distribution of the propensity scores in the matched groups before and after matching/weighting. Comparing Figure A2, (which displays propensity score distributions for the treatment and potential comparison students before matching) to Figures A3 through A7, A5 (which depict the distributions after matching) shows that matching removed most, if not all, of the differences in the propensity score distributions of the treatment and comparison groups. The literature on propensity score matching suggests that having similar propensity score distributions across the matched groups is a necessary but not sufficient condition for balance (King & Nielsen, 2016). Therefore, we also assessed to what extent matching or weighting improved the covariate balance by examining the standardized differences of each covariate between the treatment and potential comparison students prior to matching or without weights and the two groups after matching or with weights. Consistent with the sample means in Table 3, the first and sixth columns in Table 6 show that there were substantial differences between the treatment students and either of the local or statewide potential comparisons. For example, the average math and reading scores were 0.47 and 0.52 standard deviations (SDs) larger for the treatment students than the local potential comparisons. The differences were smaller for statewide comparison students but were still sizeable (0.37 and 0.45 SDs for math and reading, respectively).

The other columns in Table 6 shows that matching and weighting generally reduced these differences substantially. Columns 5 and 10 suggest that weighting yielded closely matched groups on all covariates in *both local and statewide analyses* with all differences being less than 0.02 SDs. Columns 7 through 9 suggest that all statewide matching methods also yielded closely matched groups on all covariates. For local matching, while all standardized differences were smaller than our preset threshold of 0.1 SD, some differences—especially on achievement measures—are larger than their statewide counterparts. For example, for one-to-one and radius matching, the differences for middle school test scores and passing Algebra 1 were larger than 0.05 SDs while statewide matching reduced the differences for the same measures to 0.01 SDs or less. This result is likely driven by the local matching requirement that each treatment student could only be matched with non-ECHS students from the same middle school, which may have reduced the potential set of tight matches for some treatment students with respect to baseline achievement measures. Regardless, it is important to note that even the largest differences for local methods were still within the acceptable thresholds employed by most QE applications (Stuart, 2010; What Works Clearinghouse, 2018). Next, we examine the extent to which these differences influence the bias of the QE effect estimates.

Figure 2 shows the estimated experimental benchmarks (labeled "RCT") and the local QE effect estimates (labeled to show the specific analytic method) along with their 95 percent confidence intervals (CI) for the six outcome measures. Next to each QE point estimate is the estimated bias for that estimator and its 95 percent confidence interval. Figure 3 presents the corresponding results of statewide QE analyses. Table 7 presents these results in another format, showing the point estimates and standard errors as well as the sizes of the analytic samples used in each analysis. The final column in this table shows the unadjusted effect estimates yielded by local and statewide models that do not control for any covariates, which allow us to assess the extent to which each QE method has reduced selection bias.[16] In these figures and tables, effect estimates, standard errors, and 95 percent confidence intervals are in

[16] It is easy to notice that bias of the unadjusted estimator is substantially larger than each of the QE estimators. It is also interesting that the bias of the unadjusted local estimator is larger (in absolute value) than the statewide unadjusted estimator for all outcomes.

Table 6. WSC balance statistics.

	Local					Statewide				
	Before Matching (1)	1-to-1 (2)	4-1 (3)	Radius (4)	PW (5)	Before Matching (6)	1-to-1 (7)	4-1 (8)	Radius (9)	PW (10)
Demographics										
Male	-0.24	0.01	0.00	0.00	0.00	-0.23	-0.02	0.00	-0.01	0.00
Asian	-0.02	0.08	0.04	0.05	0.00	-0.07	-0.02	0.00	-0.01	0.00
Black	0.02	-0.04	-0.02	-0.02	0.00	-0.01	-0.03	-0.01	0.01	0.00
Hispanic	0.02	0.00	0.01	0.01	0.00	-0.04	0.04	0.01	0.00	0.00
American Indian	-0.04	-0.02	0.01	0.01	0.00	-0.10	0.02	0.01	0.00	0.00
Multi-Racial	0.03	-0.01	0.01	0.01	0.00	0.04	-0.01	0.01	0.00	0.00
White	-0.03	0.03	0.00	0.00	-0.01	0.06	0.01	0.00	0.00	0.00
Free Lunch	0.04	-0.05	-0.04	-0.05	0.02	0.13	0.00	-0.01	0.01	0.01
Is ELL	-0.04	0.01	0.00	0.00	0.01	-0.07	0.04	0.01	0.00	0.00
Has Disability	-0.26	-0.02	0.00	0.01	0.00	-0.26	0.00	-0.03	-0.01	0.00
Is Gifted	0.18	-0.02	-0.04	-0.01	-0.01	0.13	0.03	0.02	0.00	0.00
Old for Grade	-0.26	-0.03	-0.01	-0.01	0.00	-0.22	0.01	0.00	0.00	0.00
MS Mobility	-0.05	-0.06	-0.05	-0.03	0.00	-0.05	-0.02	-0.01	0.00	0.00
Achievement										
MS Avg Reading Score	0.44	0.05	0.04	0.06	-0.01	0.36	0.00	0.00	0.00	-0.01
MS Avg Math Score	0.41	0.07	0.04	0.05	-0.02	0.29	0.00	0.00	0.00	-0.01
Passing Algebra I in MS	0.08	0.08	0.08	0.10	-0.02	0.02	-0.01	0.00	-0.01	-0.01
8th Grade Science Score	0.34	0.09	0.10	0.10	-0.01	0.31	0.01	0.01	0.00	-0.01
Absences										
Middle Sch Avg Days Absent	-0.21	0.01	0.00	-0.02	0.01	-0.19	0.01	-0.01	0.00	0.00

Notes: The entries in the table show the standardized differences in effect size units, which are calculated by dividing the difference between the treatment and matched comparison units by the pooled standard deviation of a given measure.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table 7. Experimental benchmarks and QE estimates.

	RCT	1-to-1	4-to-1	Radius	Prop. Weighting	OLS	Unadjusted
English 1 (matching pretest)							
Local							
Estimate	0.052	0.063	0.068	0.064	0.060	0.067	0.442
Std Error	0.025	0.017	0.014	0.013	0.015	0.020	0.032
P-Value	0.042	<0.001	<0.001	<0.001	<0.001	0.001	<0.001
Sample Size	3,385	4,078	7,729	22,983	41,857	41,857	41,857
Statewide							
Estimate	0.052	0.016	0.028	0.031	0.031	0.034	0.303
Std Error	0.025	0.016	0.015	0.014	0.014	0.017	0.054
P-Value	0.042	0.307	0.059	0.031	0.033	0.045	<0.001
Sample Size	3,385	5,816	9,976	372,351	372,758	372,758	372,758
High School Absences (matching pretest)							
Local							
Estimate	-0.160	-0.235	-0.223	-0.216	-0.213	-0.226	-0.431
Std Error	0.024	0.022	0.020	0.019	0.021	0.027	0.040
P-Value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Sample Size	3,473	4,223	8,015	24,137	45,926	45,926	45,926
Statewide							
Estimate	-0.160	-0.173	-0.166	-0.178	-0.177	-0.183	-0.325
Std Error	0.024	0.035	0.033	0.032	0.032	0.035	0.037
P-Value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Sample Size	3,473	5,986	10,239	406,402	406,810	406,810	406,810

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table 7. (Continued).

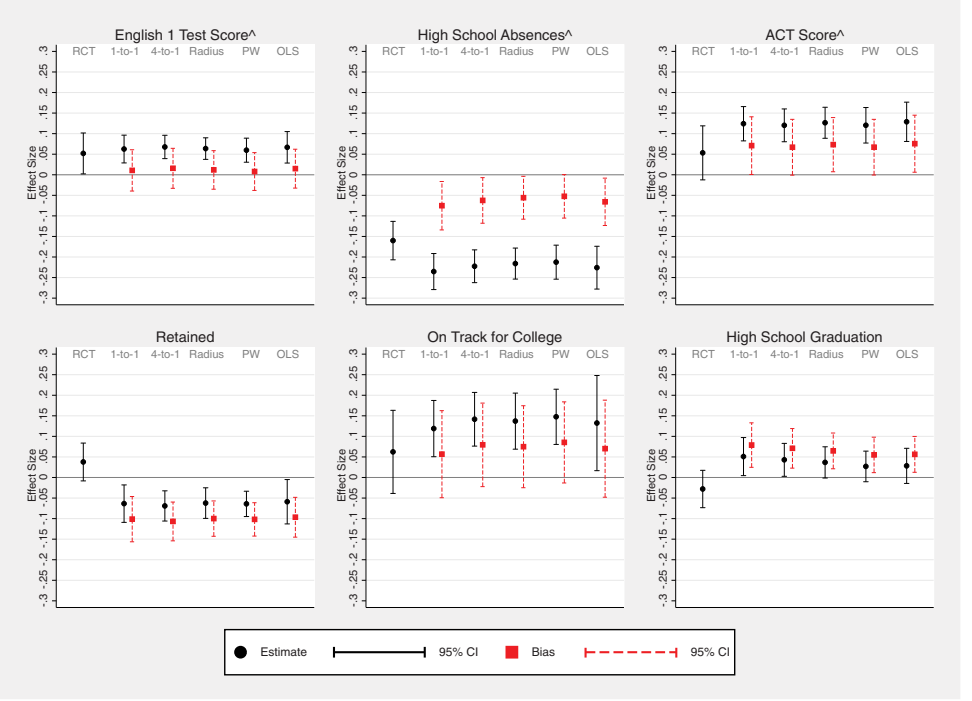
	RCT	1-to-1	4-to-1	Radius	Prop. Weighting	OLS	Unadjusted
ACT Score (matching pretest)							
Local							
Estimate	0.053	0.124	0.120	0.126	0.120	0.129	0.480
Std Error	0.034	0.021	0.020	0.019	0.022	0.024	0.038
P-Value	0.115	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Sample Size	1,985	2,314	4,354	11,920	21,505	21,505	21,505
Statewide							
Estimate	0.053	0.051	0.056	0.067	0.065	0.075	0.281
Std Error	0.034	0.029	0.028	0.026	0.026	0.029	0.068
P-Value	0.115	0.078	0.042	0.010	0.013	0.010	<0.001
Sample Size	1,985	3,570	5,728	158,347	158,548	158,548	158,548
Retained in 9th Grade							
Local							
Estimate	0.038	-0.064	-0.069	-0.062	-0.064	-0.059	-0.280
Std Error	0.023	0.023	0.019	0.019	0.016	0.028	0.035
P-Value	0.110	0.007	0.000	0.001	0.000	0.033	<0.001
Sample Size	3,394	4,121	7,806	23,330	43,486	43,486	43,486
Statewide							
Estimate	0.038	-0.031	-0.036	-0.055	-0.053	-0.048	-0.222
Std Error	0.023	0.032	0.031	0.030	0.030	0.031	0.040
P-Value	0.110	0.336	0.247	0.069	0.079	0.122	<0.001
Sample Size	3,394	5,803	9,948	385,518	385,917	385,917	385,917

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table 7. (Continued).

	RCT	1-to-1	4-to-1	Radius	Prop. Weighting	OLS	Unadjusted
				On Track for College			
Local							
Estimate	0.062	0.119	0.142	0.137	0.148	0.132	0.263
Std Error	0.052	0.035	0.033	0.035	0.034	0.059	0.061
P-Value	0.229	0.001	0.000	0.000	0.000	0.026	<0.001
Sample Size	2,878	3,421	6,446	18,675	32,667	32,667	32,667
Statewide							
Estimate	0.062	0.083	0.099	0.108	0.106	0.114	0.238
Std Error	0.052	0.045	0.042	0.041	0.041	0.049	0.052
P-Value	0.229	0.062	0.020	0.009	0.011	0.021	<0.001
Sample Size	2,878	4,796	8,204	283,834	284,192	284,192	284,192
				5-Year Graduation Rate			
Local							
Estimate	-0.028	0.051	0.043	0.037	0.027	0.028	0.279
Std Error	0.023	0.024	0.020	0.019	0.019	0.022	0.029
P-Value	0.229	0.033	0.037	0.059	0.158	0.195	<0.001
Sample Size	3,348	4,059	7,659	22,982	43,446	43,446	43,446
Statewide							
Estimate	-0.028	0.030	0.018	0.031	0.030	0.032	0.224
Std Error	0.023	0.027	0.024	0.023	0.023	0.024	0.029
P-Value	0.229	0.269	0.464	0.174	0.188	0.182	<0.001
Sample Size	3,348	5,730	9,770	384,868	385,254	385,254	385,254

Notes: This table shows RCT and the several QE estimates for each outcome. Point estimates and standard errors are in effect size units. Standard deviations used in the effect size transformation can be found in Table A1 in the Appendix.



Notes: ^ Indicates outcomes with a matching pretest. Black and red lines show 95 percent confidence intervals for effect and bias estimates.

Figure 2. Within-Study Comparison (WSC) Results—Local QE Estimates in Effect Sizes. [Color figure can be viewed at wileyonlinelibrary.com]

effect sizes units. Furthermore, Table A1 in the Appendix shows the estimated bias, its bootstrapped standard error, and the p-values of the hypothesis tests conducted for the correspondence assessment.

Table 8 shows the correspondence assessment we conducted following Steiner and Wong (2018). Recall that this framework concludes correspondence only if the equivalence test rejects the null that the difference between the two estimates is larger than the threshold (0.1 SDs in this case) *and* the difference test fails to reject that difference is zero.

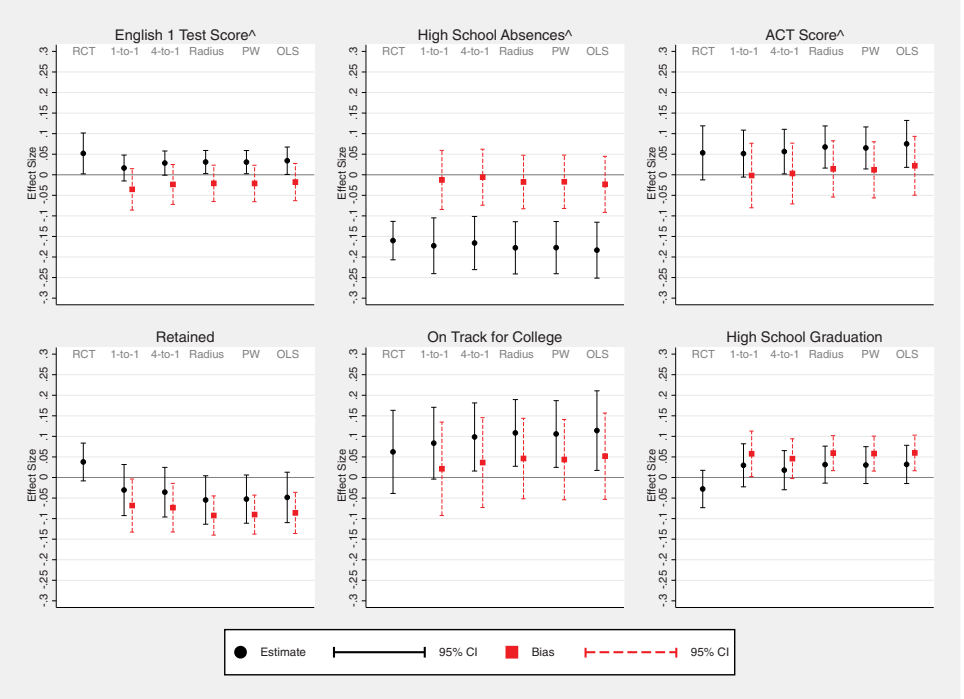
We summarize the results by examining the correspondence between the experimental benchmarks and QE estimates across the different matching and weighting methods. We start with the three outcomes we consider having natural pretests. For English 1 test scores, Table 8 suggests that all local and statewide matching and weighting methods replicated the experimental benchmarks. For two of these outcomes—absences in high school and ACT scores—all statewide QE methods replicated the experimental estimates. This is remarkably different from the results for local weighting methods, except weighting for high school absences. Table 8 shows that the local OLS model and one-to-one matching failed to replicate the benchmarks for both outcomes while local four-to-one and radius matching missed the benchmarks for absences and ACT scores, respectively. Examining the point estimates and standard errors for these two outcomes in Table 7 suggests that this result is primarily driven by substantially larger differences between the QE and experimental estimates for local models rather than for statewide models.

Table 8. Results of the correspondence assessment.

	English 1		Absences		ACT Score	
	Local	Statewide	Local	Statewide	Local	Statewide
1-to-1	Equivalence	Equivalence	Difference	Equivalence	Difference	Equivalence
4-to-1	Equivalence	Equivalence	Difference	Equivalence	Indeterminacy	Equivalence
Radius	Equivalence	Equivalence	Trivial Diff.	Equivalence	Difference	Equivalence
PW	Equivalence	Equivalence	Equivalence	Equivalence	Indeterminacy	Equivalence
OLS	Equivalence	Equivalence	Difference	Equivalence	Difference	Equivalence
	Retained		On Track		HS Graduation	
	Local	Statewide	Local	Statewide	Local	Statewide
1-to-1	Difference	Difference	Indeterminacy	Indeterminacy	Difference	Difference
4-to-1	Difference	Difference	Indeterminacy	Indeterminacy	Difference	Equivalence
Radius	Difference	Difference	Indeterminacy	Indeterminacy	Difference	Trivial Diff.
PW	Difference	Difference	Indeterminacy	Indeterminacy	Trivial Diff.	Trivial Diff.
OLS	Difference	Difference	Indeterminacy	Indeterminacy	Trivial Diff.	Trivial Diff.

Notes: This table shows the correspondence assessment results for each outcome and quasi-experimental method. PW stands for propensity weighting.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management



Notes: ^ Indicates outcomes with a matching pretest. Black and red lines show 95 percent confidence intervals for effect and bias estimates.

Figure 3. Within-Study Comparison (WSC) Results—Statewide QE Estimates in Effect Sizes [Color figure can be viewed at wileyonlinelibrary.com]

Next, we describe the results for three outcomes that lack natural pretests. For being retained in ninth grade, none of the local or statewide QE models replicated experimental results. Table 7 shows that the experimental estimate was 3.8 percent and insignificant. The local QE estimates were around -6 percent and -7 percent, and all were statistically significant. While estimates from the statewide models were somewhat closer to the experimental benchmark, all were negative, and the statistical tests do not reject that the differences are significant and larger than our 0.1 SD threshold. The direction of the bias is consistently negative (i.e., suggesting better outcomes for the early college students) for all QE models. An omitted confounder that is positively correlated with attending early college and negatively correlated with being retained, such as motivation or parental supports, can explain such negative bias in the QE estimates.

For being on track for college, the result of the correspondence assessment was indeterminacy for all QE models, indicating that we did not reject that the difference between the QE and RCT estimates is insignificant, but we rejected that the difference is smaller than the 0.1 threshold (or we did not reject the alternative hypothesis that sampling error can explain the observed difference between the two sets of estimates). Steiner and Wong (2018) argue that this may occur when the statistical tests are underpowered because of small samples or large standard errors for the experimental or QE estimates. Indeed, Table 7 shows that the standard errors for the RCT estimate and all QE estimates are much larger than the other outcomes. Table 7 also shows the statewide estimates were closer to the RCT benchmarks than

the local models but the former also had larger standard errors. This is not very surprising as this outcome does not have a natural pretest and available covariates may not do a good job of predicting it.

Finally, for high school graduation, only the four-to-one statewide matching model replicated the experimental benchmark. Local and statewide one-to-one, local four-to-one, and local radius matching methods failed to replicate the benchmark. For the remaining approaches—local and statewide weighting and OLS and statewide radius matching—the correspondence assessment yielded "trivial difference," which means that we did not reject the equivalence test (i.e., the observed difference between the two sets of estimates is trivial) but rejected that the difference is insignificant. Table 7 indicates that the differences between QE and experimental estimates were around 5 percent for the five approaches so it is reasonable to consider these differences as trivial.

To summarize, for the three outcomes with natural pretests (English 1 test scores, absences, and ACT scores), multiple QE models that replicated empirical benchmarks and for high school graduation, only one model yielded a sufficiently close QE estimate to the benchmark. For retained in ninth grade, none of the QE models replicated the experimental estimate, and for being on track for college, the imprecision of the QE estimates led to indeterminacy. It is striking that statewide models had smaller (in absolute value) biases than local models for all six outcomes (Table 7, Figures 1 and 2) and replicated the benchmarks for two outcomes (absences and ACT scores) for which local models performed poorly. Along the same lines, for four outcomes—ACT scores, being retained in ninth grade, being on track for college at the end of high school, and high school graduation—local QE estimates were positive and statistically significant while the experimental estimates were not significant, and this result does not seem to be driven by differences in the precision of effect estimates. This suggests that relying on local models for policy decisions regarding these outcomes may be misleading.

A few other observations are worth noting. Among the local models, propensity score weighting resulted in correspondence for absenteeism while the other methods did not, and it generally had smaller bias (in absolute value) than the other methods. Among the statewide models, a specific method does not stand out in terms of yielding better correspondence. Second, the direction of the QE bias is generally positive for all outcomes, i.e., QE estimates tended to favor the early colleges more than the experimental estimates. As mentioned above, this is consistent with the existence of unobserved confounders that are positively associated with attending an early college and other outcomes. Finally, local QE estimates were generally more precise than statewide estimates. This is because the cohort and feeder middle school interactions included in the local models explain a considerable proportion of the outcome variance that is not explained by the other covariates.

DISCUSSION

Existing WSCs in education highlight that a matching pretest is the most important covariate for minimizing QE bias. In addition, there are very few education WSCs that examine interventions targeting high school and postsecondary students; for example, none of the 12 WSCs included in the Wong, Valentine, and Miller-Bain (2017) review evaluated a high school intervention. QE studies of high school interventions that aim to boost students' access to postsecondary education are challenging because many key outcomes do not have natural pretests at the student level (e.g., high school graduation, being academically prepared for college). Therefore, this paper makes an important contribution to the education WSCs concerning evaluations of high school and postsecondary education interventions. We conclude that

researchers examining a similar high school intervention with a similar selection mechanism can expect to produce QE results with minimal bias for outcomes with natural pretests using variables that are typically available in extant longitudinal databases. This conclusion is generally consistent with the existing education WSCs that examined elementary and middle-grade interventions. For outcomes without a natural pretest, however, our results suggest that the researchers need to be cautious and the covariates typically found in administrative data sets may not adequately capture all potential confounders. It would be interesting to see whether supporting the covariates used in our analyses with additional covariates at the student or school level would decrease or eliminate the QE bias. For example, one could create additional middle school-level covariates for prior cohorts of students including school climate, participation in dual-enrollment in high school, and college enrollment post high school. While local models may control for these factors implicitly, this may be offset by the locational restriction they place on the comparison groups.

Indeed, one of the striking results of our analyses is that local models generally failed to replicate experimental benchmarks. This was true even for two of the three outcomes with natural pretests. This result can partially be explained because local matching and weighting methods did not yield well-balanced treatment and comparison groups in this WSC. However, given that the treatment-comparison group differences were smaller than the conventionally used thresholds and statewide models replicated benchmarks for at least some outcomes, it is likely that, compared to the statewide approach, the geographical restriction imposed by the local approach yielded inferior comparison groups. The implications of this for applied researchers are that local restrictions may do more harm than good if good local matches are small in number. It is possible that prioritizing balance on observable covariates among a relatively small number of good matches may have distorted balance on unobservable confounders such as motivation (Wong, Valentine, & Miller-Bain, 2017). A promising avenue for future research is to combine local and non-local matches as suggested by Stuart and Rubin (2008).

Another important result is that QE bias did not vary by how the estimated propensity scores were used in the analysis, which was especially relevant for statewide analyses. This may be a direct result of the doubly robust approach as all QE methods utilized the same set of covariates. Related to this observation is that among the statewide approaches, the OLS approach (which used all available potential comparison students with equal weights) yielded generally similar coefficient estimates and standard errors to the other approaches. This creates some ambiguity about the need to conduct matching and weighting as a data preprocessing step prior to analysis.

A limitation of this study is that it conceives of the selection problem as one of individual choice rather than institutional constraints or facilitation. Future work should examine whether individual and school predictors may combine to push students into early colleges or other high school interventions. An important methodological step that would make this investigation more feasible is a recent approach to match on an optimal mix of student and school factors to achieve good balance on observable baseline covariates (Pimentel et al., 2018; Zubizareeta & Keele, 2017).

Finally, this paper showed the promises and pitfalls of replication with a large sample and many covariates. Future research could investigate more limiting cases. For example, how small does the pool of potential comparison cases need to get before it is extremely unlikely to replicate the RCT result? Or, in models with matching pretests, what is the minimum set of focal covariates necessary to replicate the RCT result?

FATIH UNLU is a Senior Economist at the RAND Corporation, 1776 Main Street, Santa Monica, CA 90401 (e-mail: funlu@rand.org).

DOUGLAS LEE LAUEN is a Professor of Public Policy and Sociology at the University of North Carolina at Chapel Hill, Department of Public Policy, UNC–Chapel Hill, Abernethy Hall, CB#3435, Room 121A, Chapel Hill, NC 27599-3435 (e-mail: dlauen@unc.edu).

SARAH CRITTENDEN FULLER is a Research Associate Professor at the University of North Carolina at Chapel Hill, Education Policy Initiative at Carolina, University of North Carolina, 314 Cloister Court, Chapel Hill, NC 27514 (e-mail: sarah.fuller@unc.edu).

TIFFANY BERGLUND is a Quantitative Analyst III at the RAND Corporation, 1776 Main Street, Santa Monica, CA 90401 (e-mail: ttsai@rand.org).

ELC ESTRERA is a Ph.D. Candidate at the University of North Carolina at Chapel Hill, UNC–Chapel Hill, Abernethy Hall, CB#3435, Chapel Hill, NC 27599-3435 (e-mail: estrera@live.unc.edu).

ACKNOWLEDGMENTS

This project is funded by a grant from the Institute of Education Sciences, U.S. Department of Education (grant number: R305A150477). The opinions expressed here are those of the authors and do not represent views of the Institute or the Department of Education. We thank Brian Phillips, Jane Furey, Josh Horvath, and Anna Rybinska for excellent research assistance. We also thank Alisa Chapman, Tom Cook, Julie Edmunds, and Elizabeth Stuart for their contributions to this project, the editors of JPAM, and three excellent anonymous reviewers. We also gratefully acknowledge the North Carolina Department of Public Instruction, the North Carolina Community College System, and the University of North Carolina System Office for supporting this project with data and staff resources. Part of the research was conducted when Unlu was at Abt Associates.

REFERENCES

Abdulkadiroglu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. The Quarterly Journal of Economics, 126, 699–748.

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. Review of Economics and Statistics, 86, 180–194.

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

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61, 962–973.

Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. Journal of Policy Analysis and Management, 31, 729–751.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22, 31–72.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. Journal of Policy Analysis and Management 37, 403–429.

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. American Economic Review, 104, 2633–2679.

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

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of the pretest as a covariate, unreliable measurement and mode of data analysis. Psychological Methods, 15, 56–68.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). Assessing how bias reduction is influenced by covariate choice, unreliability and data analytic mode: An analysis of different kinds of within-study comparisons in different substantive domains. Multivariate Behavioral Research, 44, 828–847.

Cook, T. D., Zhu, N., Klein, A., Starkey, P., & Thomas, J. (2020). How much bias results if a quasi-experimental design combines local comparison groups, a pretest outcome measure and other covariates?: A within study comparison of preschool effects. Psychological Methods, 25(6), 726–746.

Decker, P. T. (2014). Presidential address: False choices, policy framing, and the promise of "Big Data." Journal of Policy Analysis and Management, 33, 252–262.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94, 1053–1062.

Dong, N., & Lipsey, M. W. (2018). Can propensity score analysis approximate experiments using pretest and demographic information in Pre-K intervention research? Evaluation Review, 42, 34–70.

Edmunds, J. A., Bernstein, L., Glennie, E., Willse, J., Arshavsky, N., Unlu, F., … Dallas, A. (2010). Preparing students for college: The implementation and impact of the early college high school model. Peabody Journal of Education, 85, 348–364.

Edmunds, J. A., Bernstein, L., Unlu, F., Glennie, E., Willse, J., Smith, A., & Arshavsky, N. (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, 136–159.

Edmunds, J. A., Unlu, F., Furey, J., Glennie, E., & Arshavsky, N. (2020). What happens when you combine high school and college? The impact of the early college model on postsecondary performance and completion. Educational Evaluation and Policy Analysis, 42, 257–278.

Edmunds, J. A., Unlu, F., Glennie, E., Bernstein, L., Fesler, L., Furey, J., & Arshavsky, N. (2017). Smoothing the transition to postsecondary education: The impact of the early college model. Journal of Research on Educational Effectiveness, 10, 297–325.

Edmunds, J. A., Willse, J., Arshavsky, N., & Dallas, A. (2013). Mandated engagement: The impact of early college high schools. Teachers College Record, 115, 31.

Fortson, K., Verbitsky-Savitz, N., Kopa, E., & Gleason, P. (2012). Using an experimental evaluation of charter schools to test whether nonexperimental comparison group methods can replicate experimental impact estimates. NCEE, 2012–4019.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. Journal of Human Resources, 22, 194–227.

Furgeson, J., Gill, B., Haimson, J., Killewald, A., McCullough, M., Nichols-Barrer, I., Teh, B.-R., … Lake, R. (2012). Charter-school management organizations: Diverse strategies and diverse student impacts. Princeton, NJ: Mathematica Policy Research, Inc.

Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for constructing and assessing propensity scores. Health Services Research, 49, 1701–1720.

Gill, B., Furgeson, B., Chiang, H., Teh, B.-R., Haimson J., & Savitz, N. V. (2016) Replicating experimental impact estimates with nonexperimental methods in the context of control-group noncompliance. Statistics and Public Policy, 3, 1–11.

Hallberg, K., Cook, T. D., Steiner, P. M., & Clark, M. H. (2018). Pretest measures of the study outcome and the elimination of selection bias: Evidence from three within-study comparisons. Prevention Science, 19, 274–283.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The Review of Economic Studies, 64, 605–654.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. The Review of Economic Studies, 65, 261–294.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. In A. Gelman & X. L. Meng (Eds.), Applied Bayesian and causal inference from an incomplete data perspective (pp. 49–60). New York, NY: Wiley.

Katz, L. G. (2000). Academic redshirting and young children. ERIC Digest EDO-PS-00-13.

King, G., & Nielsen, R. (2016). Why propensity scores should not be used for matching. Harvard University, Working Paper.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 76, 604–620.

McDonald, D., & Farrell, T. (2012). Out of the mouths of babes: Early college high school students' transformational learning experiences. Journal of Advanced Academics, 23, 217–248.

Pimentel, S. D., Page, L. C., Lenard, M., & Keele, L. (2018). Optimal multilevel matching using network flows: An application to a summer reading intervention. The Annals of Applied Statistics, 12, 1479–1505.

Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X., & Lai, S. (2010). Nonparametric bootstrapping for hierarchical data. Journal of Applied Statistics, 37, 1487–1498.

Roderick, M., Nagaoka, J., & Coca, V. (2009). College readiness for all: The challenge for urban high schools. The Future of Children, 19, 185–210.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 39, 33–38.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton-Mifflin.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, 125, 305–353.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, H. M. (2010). The importance of covariate selection in controlling for selection bias in observational studies. Psychological Methods, 15, 250–267.

Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. Evaluation Review, 42, 214–247.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 25, 1.

Stuart, E. A., & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. Journal of Educational and Behavioral Statistics, 33, 279–306.

Tipton, E., & Olsen, R. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. Educational Researcher, 47(8).

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). KIPP middle schools: Impacts on achievement and other outcomes. Princeton, NJ: Mathematica Policy Research.

U.S. Department of Education, Institute of Education Sciences. (2018). Washington, DC: National Center for Education Evaluation and Regional Assistance Report (NCEE 2018–4013).

What Works Clearinghouse. (2018). Standards Handbook Version 4.0. Available at https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.

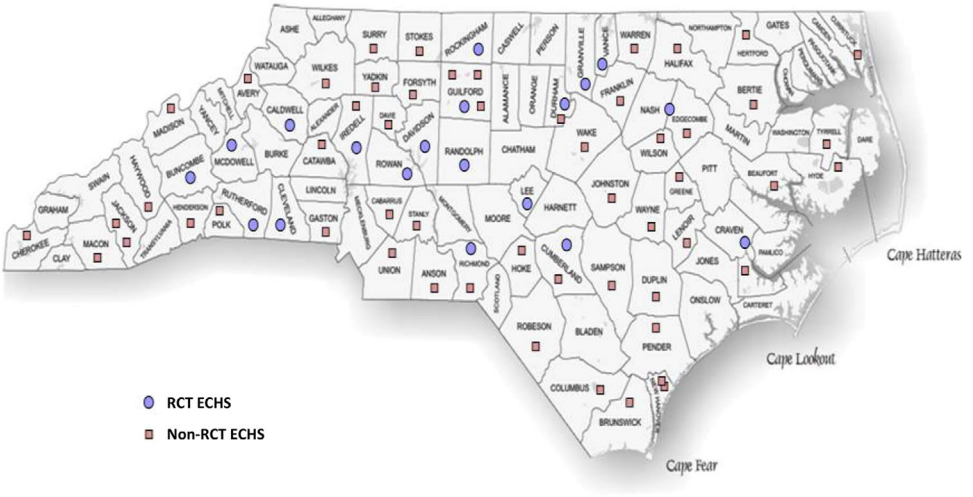
Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. Journal of Policy Analysis and Management, 26, 455–477.

Wong, V. C., & Steiner, P. M. (2018). Designs of empirical evaluations of non-experimental methods in field settings. Evaluation Review, 42, 176–213.

Wong, V. C., Valentine, J., & Miller-Bain, K. (2017). Covariate selection in education observation studies: A review of results from within-study comparisons. Journal on Research on Educational Effectiveness, 10, 207–236.

Zubizarreta, J. R., & Keele, L. (2017). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. Journal of the American Statistical Association, 112, 547–560.

APPENDIX



Note: RCT ECHS sites are those Early Colleges that have at least one cohort in the experiment used in the WSC.

Figure A1. Locations of Early College High Schools (ECHS) in North Carolina. [Color figure can be viewed at wileyonlinelibrary.com]

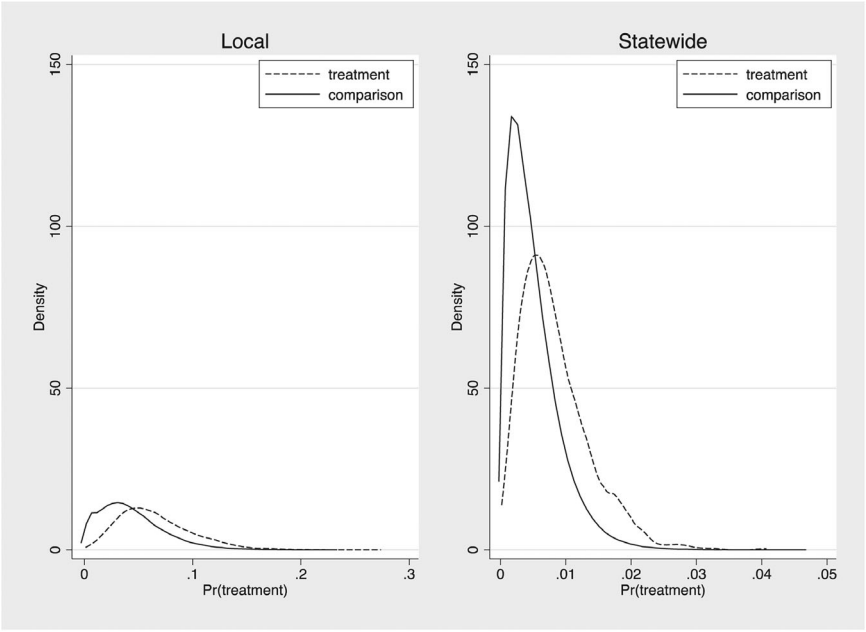


Figure A2. Propensity Score Distributions Before Matching, Local and Statewide Models.

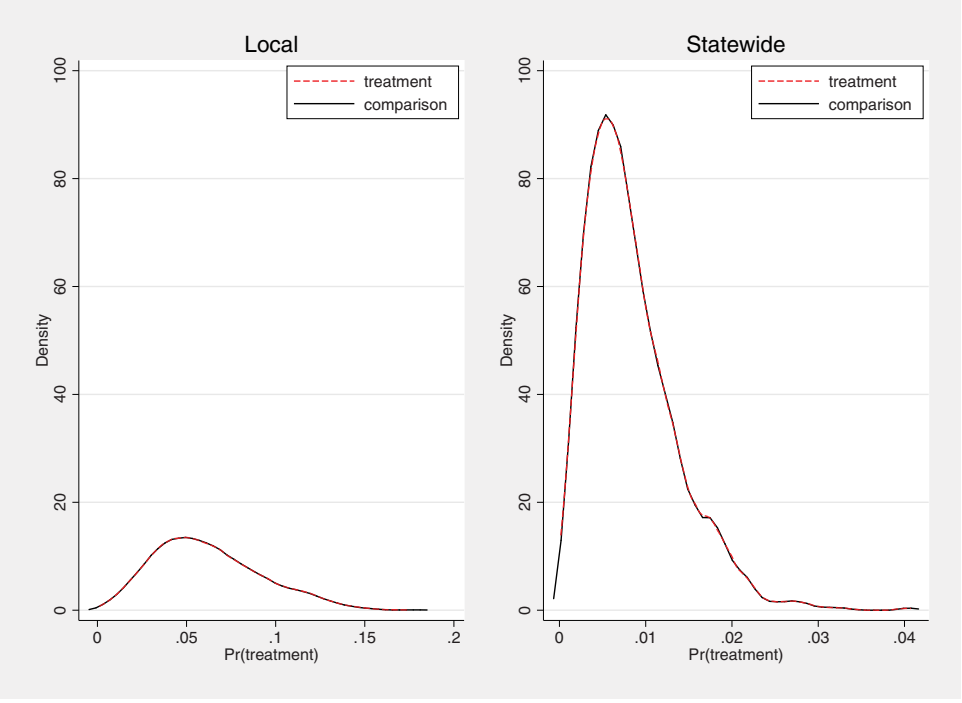


Figure A3. Propensity Score Distributions with 1-to-1 Matching, Local and Statewide. [Color figure can be viewed at wileyonlinelibrary.com]

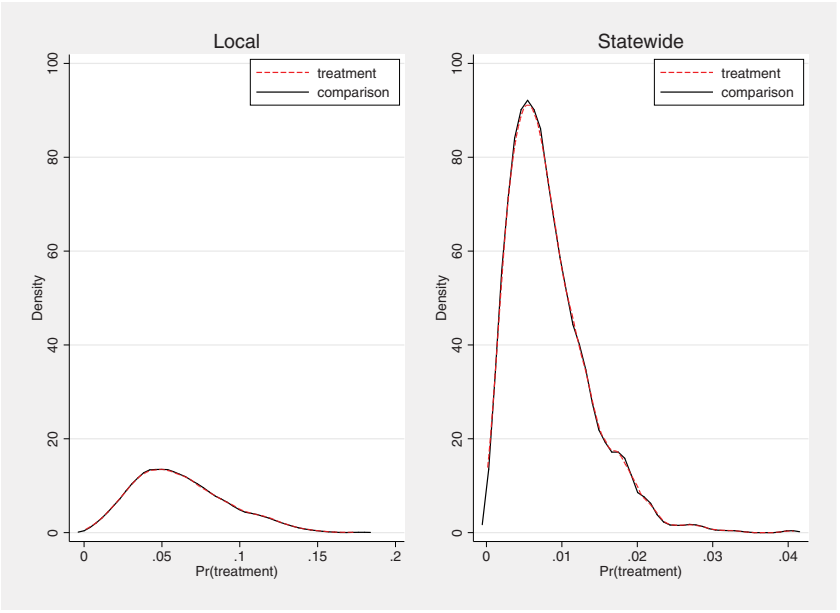


Figure A4. Propensity Score Distributions with 4-to-1 Matching, Local and Statewide. [Color figure can be viewed at wileyonlinelibrary.com]

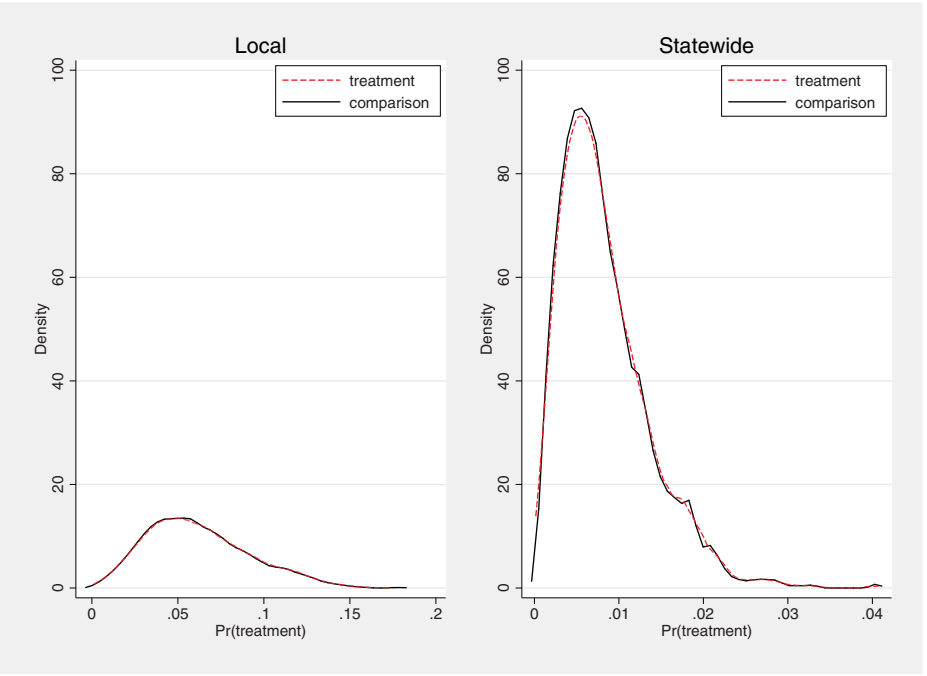


Figure A5. Propensity Score Distributions with Radius Matching, Local and Statewide. [Color figure can be viewed at wileyonlinelibrary.com]

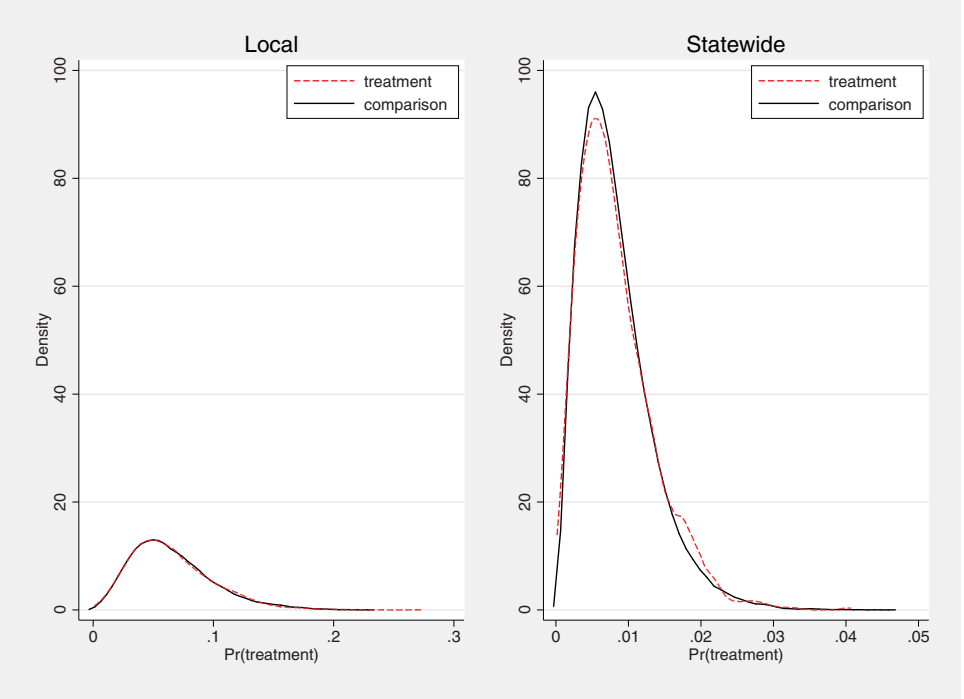


Figure A6. Propensity Score Distributions with Propensity Score Weighting, Local and Statewide. [Color figure can be viewed at wileyonlinelibrary.com]

Table A1. Quasi-experimental bias estimates and correspondence test results.

	1-to-1	4-to-1	Radius	PW	OLS
	English 1 (matching pretest)				
SD Outcome	0.968	0.968	0.968	0.968	0.968
Local					
Bias	0.011	0.016	0.012	0.008	0.015
Std Error	0.026	0.025	0.024	0.023	0.024
P-value H_0^d	0.678	0.527	0.622	0.737	0.538
P-value H_e^{01}	<0.001	<0.001	<0.001	<0.001	<0.001
P-value H_{02}^e	<0.001	<0.001	<0.001	<0.001	<0.001
Statewide					
Bias	-0.035	-0.024	-0.021	-0.021	-0.018
Std Error	0.026	0.025	0.023	0.023	0.023
P-value H_0^d	0.167	0.342	0.354	0.349	0.443
P-value H_{01}^e	<0.001	<0.001	<0.001	<0.001	<0.001
P-value H_{02}^e	0.006	0.001	<0.001	<0.001	<0.001
	High School Absences (matching pretest)				
Local					
SD Outcome	10.852	10.852	10.852	10.852	10.852
Bias	-0.075	-0.062	-0.056	-0.052	-0.066
Std Error	0.030	0.028	0.027	0.027	0.029
P-value H_0^d	0.012	0.027	0.036	0.052	0.025
P-value H_{01}^e	<0.001	<0.001	<0.001	<0.001	<0.001
P-value H_{02}^e	0.205	0.092	0.049	0.039	0.122
Statewide					
SD Outcome	10.852	10.852	10.852	10.852	10.852
Bias	-0.013	-0.006	-0.018	-0.017	-0.023
Std Error	0.037	0.035	0.033	0.033	0.035
P-value H_0^d	0.732	0.865	0.596	0.605	0.503
P-value H_{01}^e	0.001	0.001	<0.001	<0.001	<0.001
P-value H_{02}^e	0.009	0.003	0.007	0.006	0.014
	ACT Score (matching pretest)				
Local					
SD Outcome	5.055	5.055	5.055	5.055	5.055
Bias	0.071	0.067	0.073	0.067	0.076
Std Error	0.036	0.035	0.034	0.035	0.035
P-value H_0^d	0.048	0.053	0.029	0.052	0.033
P-value H_{01}^e	0.208	0.170	0.213	0.170	0.245
P-value H_{02}^e	<0.001	<0.001	<0.001	<0.001	<0.001
Statewide					
SD Outcome	5.055	5.055	5.055	5.055	5.055
Bias	-0.002	0.003	0.014	0.012	0.022
Std Error	0.040	0.038	0.035	0.035	0.036
P-value H_0^d	0.964	0.935	0.684	0.732	0.549
P-value H_{01}^e	0.006	0.005	0.007	0.006	0.016
P-value H_{02}^e	0.007	0.003	0.001	0.001	<0.001

Table A1. (Continued).

	1-to-1	4-to-1	Radius	Prop. Weighting	OLS
Retained in 9th Grade					
Local					
SD Outcome	0.305	0.305	0.305	0.305	0.305
Bias	-0.101	-0.107	-0.100	-0.102	-0.097
Std Error	0.028	0.024	0.022	0.021	0.025
P-value H_0^d	<0.001	<0.001	<0.001	<0.001	<0.001
P-value H_{01}^e	<0.001	<0.001	<0.001	<0.001	<0.001
P-value $H_0^e 2$	0.519	0.613	0.500	0.536	0.447
Statewide					
SD Outcome	0.305	0.305	0.305	0.305	0.305
Bias	-0.068	-0.073	-0.092	-0.090	-0.086
Std Error	0.033	0.030	0.024	0.024	0.026
P-value H_0^d	0.038	0.015	<0.001	<0.001	0.001
P-value H_{01}^e	<0.001	<0.001	<0.001	<0.001	<0.001
P-value H_{02}^e	0.167	0.189	0.378	0.344	0.294
On Track for College					
Local					
SD Outcome	0.452	0.452	0.452	0.452	0.452
Bias	0.057	0.079	0.075	0.085	0.070
Std Error	0.054	0.052	0.051	0.050	0.060
P-value H_0^d	0.294	0.126	0.141	0.090	0.245
P-value H_{01}^e	0.211	0.345	0.310	0.385	0.310
P-value H_{02}^e	0.002	<0.001	<0.001	<0.001	0.002
Statewide					
SD Outcome	0.452	0.452	0.452	0.452	0.452
Bias	0.021	0.036	0.046	0.044	0.052
Std Error	0.058	0.056	0.050	0.050	0.054
P-value H_0^d	0.716	0.516	0.355	0.383	0.333
P-value H_{01}^e	0.087	0.126	0.140	0.129	0.184
P-value H_{02}^e	0.018	0.007	0.002	0.002	0.002
5-Year Graduation Rate					
Local					
SD Outcome	0.405	0.405	0.405	0.405	0.405
Bias	0.079	0.071	0.065	0.055	0.056
Std Error	0.028	0.025	0.022	0.022	0.022
P-value H_0^d	0.004	0.004	0.004	0.013	0.011
P-value H_{01}^e	0.221	0.118	0.056	0.020	0.025
P-value H_{02}^e	<0.001	<0.001	<0.001	<0.001	<0.001
Statewide					
SD Outcome	0.405	0.405	0.405	0.405	0.405
Bias	0.058	0.046	0.059	0.058	0.060
Std Error	0.028	0.025	0.022	0.022	0.022
P-value H_0^d	0.042	0.064	0.006	0.007	0.007
P-value H_{01}^e	0.067	0.014	0.030	0.027	0.034
P-value H_{02}^e	<0.001	<0.001	<0.001	<0.001	<0.001

Notes: This table shows the estimated bias and its bootstrapped standard error (in effect sizes) for each QE method. "$P-value\, H_0^d$" shows the p-value of the null hypothesis that states the bias is zero ($H_0^d: \ QE\ Bias\ = \ 0$ "$P-value\, H_{01}^e$" and "$P-value\, H_{02}^e$" are p-values from two one-sided "equivalence" tests with the following null hypotheses: $H_{01}^e: \ QE\ Bias \geq \delta_E$ and $H_{02}^e: \ QE\ Bias \leq -\delta_E \cdot \delta_E$ is set to 0.1 standard deviations. Please see the text for a more detailed description of these tests.