Title: Can Quasi-Experimental Evaluations that Rely on State Longitudinal Data Systems Replicate Experimental Results: Findings from a Within-Study Comparison

Authors:
Fatih Unlu, RAND Corporation
Douglas Lee Lauen, University of North Carolina Chapel Hill
Sarah Crittenden Fuller, University of North Carolina Chapel Hill
Tiffany Tsai, RAND Corporation

Background/Context: This paper addresses whether researchers should expect to obtain unbiased effect estimates from quasi-experimental (QE) studies conducted with baseline covariates typically available in longitudinal administrative state databases. Specifically, we report findings from a within-study comparison (WSC) combining data from a randomized control trial (RCT) evaluating early college high schools (ECHS) in North Carolina with administrative data including pre- and post-treatment information on potential comparison students who did not participate in the RCT.

Purpose/Objective/Research Question: We examine a high school intervention aimed at boosting access to postsecondary education. Conducting QE studies for such interventions is challenging as most commonly used outcome measures do not have natural student-level pretests (e.g., high school graduation, academic preparation for college). Existing WSCs in education highlight pretests as the most important covariate for minimizing QE bias. Furthermore, most education WSCs examine interventions targeting elementary/middle grades or postsecondary students; for example, none of the 12 WSCs included in Wong et al. (2017) evaluated a high school intervention. Therefore, we make an important contribution to the literature on education WSCs concerning high school and postsecondary interventions that examine outcome measures without natural pretests.

Setting: Nationally, there are over 240 ECHSs in 28 states. North Carolina is home to more than 80, which is approximately thirty percent of all ECHSs in the nation – more than any other state.

Population/Participants/Subjects: The experimental sample includes six cohorts of high school students who applied to enroll in one of the 19 early colleges included in the RCT and participated in lotteries between the 2004-05 and 2010-11 school years (2,174 treatment and 1,584 control). The QE estimates are obtained using more than 600,000 students from the same 9th grade cohorts who did not participate in these lotteries and did not enroll in ECHSs ("potential comparisons"). Table 1 provides an overview of these cohorts and Table 2 presents descriptive statistics for the experimental and QE analytic samples.

Intervention: ECHSs are small schools (between 100 and 400 students) primarily located on campuses of two- or four-year colleges. ECHS students earn, at no financial cost to them, up to two years of transferable college credit or an associate's degree while simultaneously satisfying state high school graduation requirements.

Data Sources/Measures: We use a rich longitudinal student-level data set constructed using administrative data from the North Carolina Department of Public Instruction. We focus on 6 high school outcomes[1] – English I test scores, 9th grade retention, average attendance through high school, five-year high school graduation, ACT test scores, and being on track for college in 12th grade. We also have many variables measured prior to entry into high school that are used to control for confounding, including demographic variables; prior middle school performance including 6th to 8th grade math and reading test scores, 8th grade science test scores, and passing Algebra I in middle school; middle school attendance; and mobility during middle school.

[1] We are in the process of conducting the described WSC analyses for additional outcomes (e.g., postsecondary enrollment during and post high school, attainment of postsecondary degrees), which will be completed and incorporated to our presentation if our proposal is accepted.

Analysis: We compared RCT impact estimates (experimental benchmarks) to estimates from eight QE estimators that utilized the treatment group from the RCT and a variety of comparison groups. The QE estimators differed by how the comparison groups were constructed but all used the same set of covariates described above. Four QE estimators restricted the comparison group to non-ECHS students who attended the same middle schools as the treatment students ("local comparisons"). The purpose of this restriction was to account for historical and locational factors may not be fully captured by the student-level controls. Three of the local estimators utilized propensity scoring techniques (1-to-1 matching, radius matching, and inverse propensity score weighting or IPW) and the last local estimator included all local non-ECHS students and controlled for covariates in a regression model (OLS). The remaining four QE estimators (1-to-1 matching, radius matching, IPW, and OLS) placed no locational restrictions on the comparison group ("statewide comparisons"). All QE estimators were estimated using regression models that controlled for all matching covariates ("doubly robust"). Table 3 provides more details on these estimators.

We assessed correspondence between benchmarks and QE estimates using the correspondence framework of Steiner and Wong (2018). This entailed formally testing the insignificance of the difference between two estimates (which accounted for the correlation between two estimates using bootstrapping) and statistical equivalence of the two estimates (carried out using two one-sided hypothesis tests that assessed the difference between the two estimates being smaller than 0.10 standard deviations). Figure 1 presents four potential outcomes of this assessment (equivalence, trivial difference, indeterminacy, and difference).

Findings: Table 4 shows that baseline differences between the treatment and all potential comparison students were sizeable and propensity scoring methods substantially reduced these differences. Figures 2 and 3 show the experimental benchmarks and the local and statewide QE estimates for the six outcomes respectively. Figure 4 shows the results of the correspondence assessment for the eight QE estimators:

- For three outcomes with natural/proxy pretests (English 1 scores, absences, and ACT scores), all statewide QE estimators replicated the experimental findings. The performance of the local estimators were mixed: all local estimators replicated benchmarks for English 1 scores and only the OLS estimator did so for ACT scores.
- For high school graduation, local 1-to-1 estimator replicated the benchmark and the RCT-QE differences were trivial for other estimators.
- For retained in 9th grade, all QE estimators failed to replicate the benchmark.
- For on-track, the assessment result was "indeterminacy" for all QE estimators (except OLS) which is potentially due to imprecise point estimates.

Conclusions:

- Natural/proxy pretest is critical for reducing QE bias (consistent with education WSCs);
- Imposing locational restrictions on comparison groups does not perform better (and performs worse in some cases) than QE estimators with no such restrictions (inconsistent with some education WSCs);

- The QE bias is generally insensitive to the specific QE analytic method used, such as 1-to-1 matching or OLS (consistent with education WSCs).

References

Steiner, P. M., & Wong, V. C. (2018). Assessing Correspondence Between Experimental and Nonexperimental Estimates in Within-Study Comparisons. *Evaluation Review*, *42*(2), 214–247. https://doi.org/10.1177/0193841X18773807

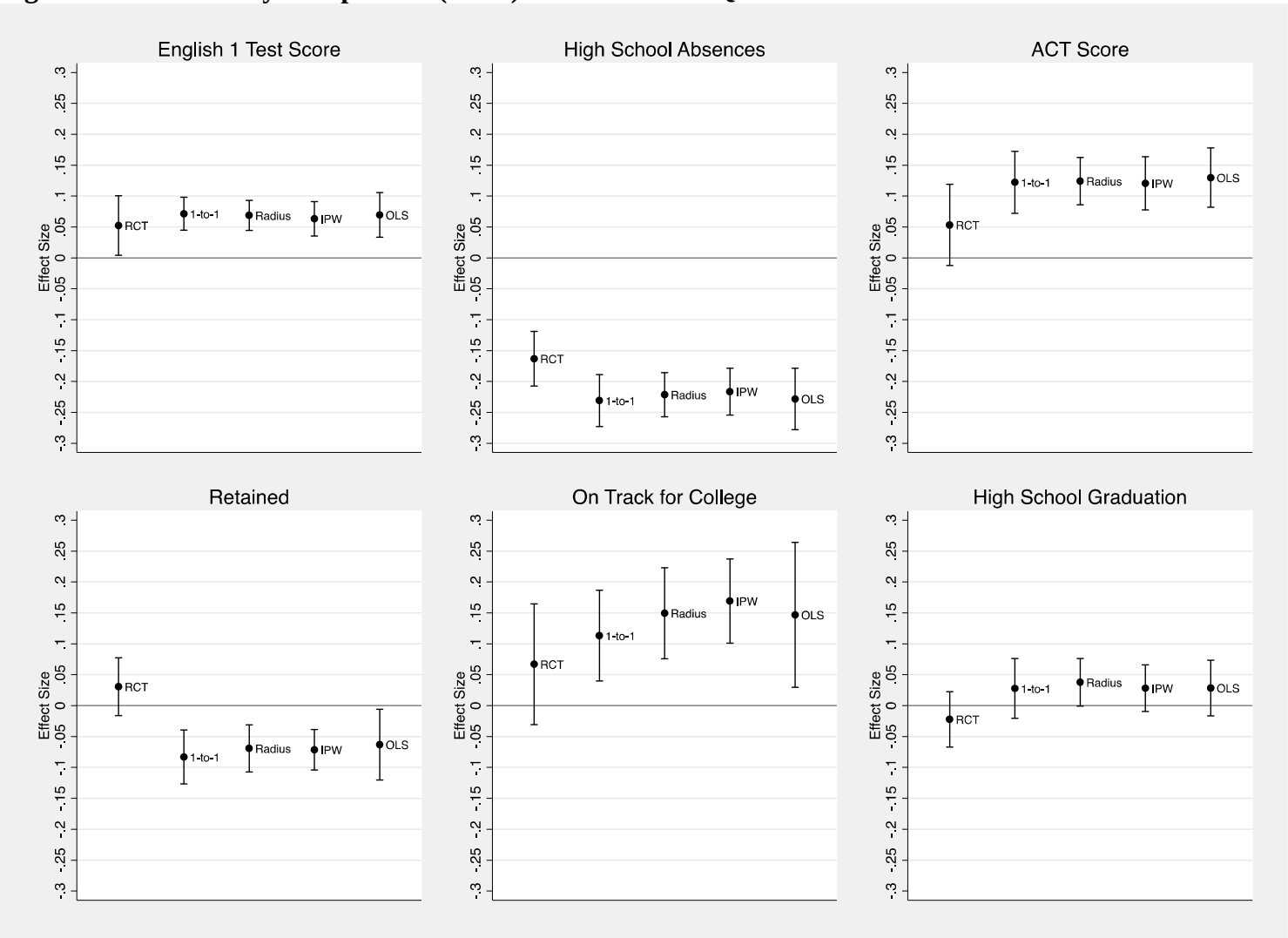
Wong, V.C., Valentine, J., Miller-Bain, K. (2017). Covariate Selection in Education Observation Studies: A Review of Results from Within-study Comparisons. Journal on Research on Educational Effectiveness, 10 (1), 207-236.

Figure 1. Four Outcomes of the Correspondence Test by Steiner and Wong (2018)

	Equivalence (C^E)	
Insignificant Difference (C^D)	$C^E = 0$: Noncorrespondence (Insignificant Equivalence)	$C^E = 1$: Correspondence (Significant Equivalence)
$C^D = 0$: Noncorrespondence (significant difference)	Difference	Trivial difference
$C^D = 1$: Correspondence (insignificant difference)	Indeterminacy	Equivalence

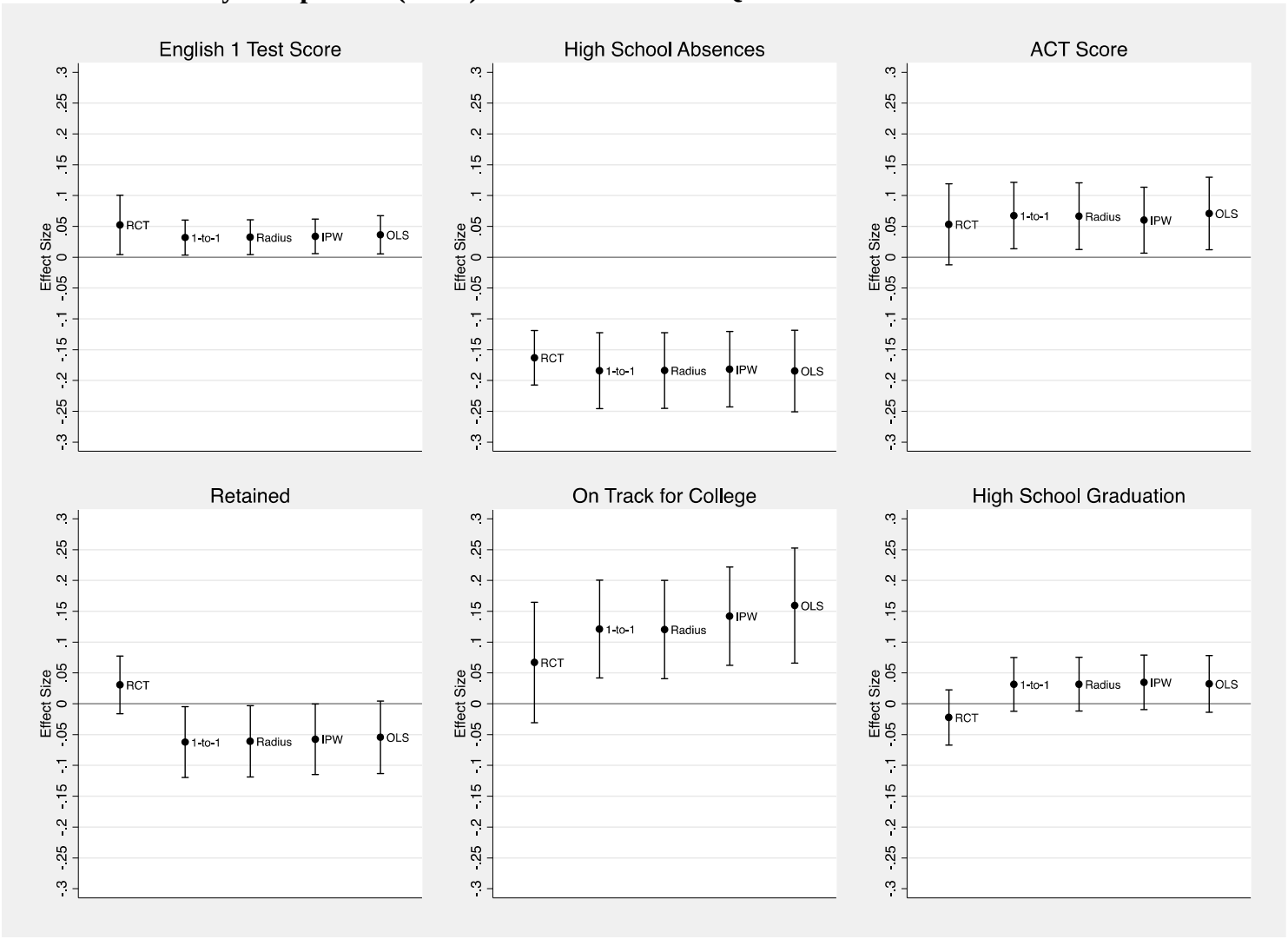
Source: Table 1 in Steiner and Wong (2018).

Figure 2. Within-Study Comparison (WSC) Results – Local QE Estimates in Effect Sizes



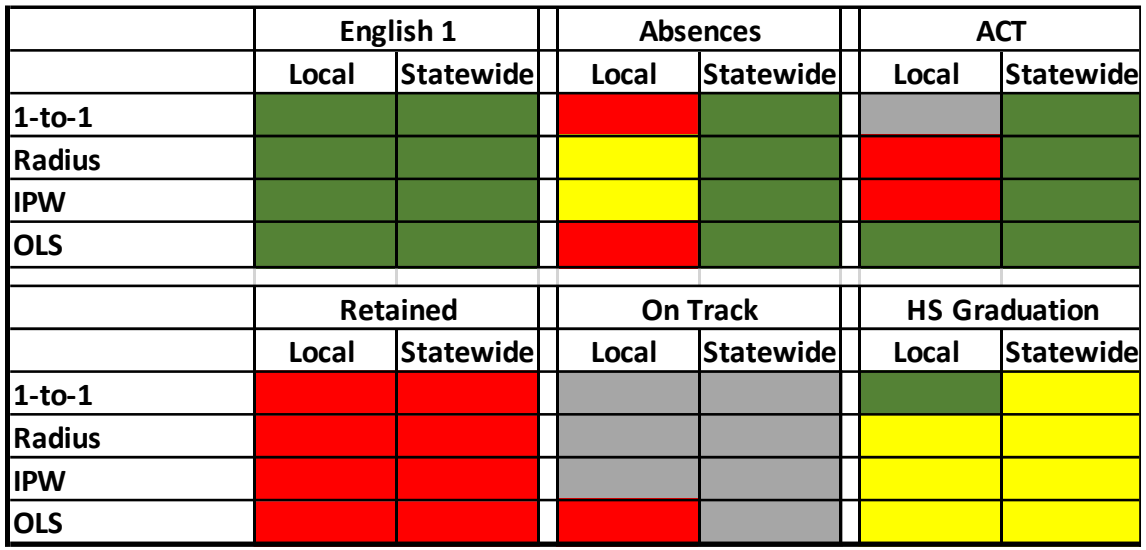
Note: Each figure shows the RCT benchmark and the QE estimates from four local models (1-1 matching, radius matching, IPW, and OLS). Point estimates and 95 % confidence intervals are displayed in effect size units (which are calculated by dividing the effect estimates and confidence interval boundaries by the pooled standard deviation of the outcome).

Figure 3. Within-Study Comparison (WSC) Results – Statewide QE Estimates in Effect Sizes



Note: Each figure shows the RCT benchmark and the QE estimates from four statewide models (1-1 matching, radius matching, IPW, and OLS). Point estimates and 95 % confidence intervals are displayed in effect size units (which are calculated by dividing the effect estimates and confidence interval boundaries by the pooled standard deviation of the outcome).

Figure 3. Correspondence between RCT Benchmarks and QE Estimates



Notes: Green cells denote "equivalance", yellow cells denote "trivial difference", gray cells denote "indeterminacy", and red cells denote "difference".

Table 1. Cohorts Included in Study

	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14
8th grade	1	2	3	4	5	6				
9th grade		1	2	3	4	5	6			
10th grade			1	2	3	4	5	6		
11th grade				1	2	3	4	5	6	
12th grade					1	2	3	4	5	6
Postsecondary/13th grade						1	2	3	4	5
Postsecondary/14th grade							1	2	3	4
Postsecondary/15th grade								1	2	3

Table 2. Means of Covariates for WSC Samples

	Treatment Group	Potential Local Comparisons	Potential Statewide Comparisons
Demographics			
Male	40%	52%	51%
Asian	1%	1%	2%
Black	28%	27%	29%
Hispanic	8%	7%	8%
American Indian	0%	1%	2%
Multi Racial	4%	3%	3%
White	59%	60%	56%
Free Lunch	52%	50%	45%
Is LEP	3%	4%	5%
Has Disability Status	4%	13%	13%
Is Gifted	21%	15%	16%
Old for Grade	12%	22%	20%
MS Mobility	23%	25%	20%
Achievement			
Passed Alg 1 in MS	22%	19%	18%
Middle Sch. Avg Math Scr. (z-score)	0.25	-0.12	-0.01
Middle Sch. Avg Reading Sc. (z-score)	0.31	-0.09	-0.01
Grade 8 Science Score (z-score)	0.18	-0.11	-0.05
Absences			
Middle School Avg Days Absent	6.58	8.05	7.93
Number of Observations	2174	46648	615683

Notes: Middle school average test scores and days absent are simple averages of the same measures in the 6th, 7th, and 8th grades. A student could be old for grade if he or she was retained in a prior grade or because of kindergarten redshirting.

Table 3. Quasi-Experimental Models

Label	Location Restriction for Potential Comparisons	Propensity Score Estimation	Details on Matching	Additional Controls
Local OLS	Non-ECHS students from same feeder middle schools as treatment students	N.A	N.A.	Cohort by feeder middle school interactions
Local 1-1 Matching		Probit	1-1*	
Local Radius Matching			Radius*	
Local IPW			N.A.	
Statewide OLS	All non-ECHS students in NC	N.A	N.A.	-
Statewide 1-1 Matching		Probit	1-1	
Statewide Radius Matching			Radius	
Statewide IPW			N.A.	

Notes: *Local 1-1 and radius matching estimators implemented exact matching on cohort and feeder middle schools.

Table 4. WSC Balance Statistics

	Local				Statewide			
	Before Matching	1-to-1	Radius	IPW	Before Matching	1-to-1	Radius	IPW
Demographics								
Male	-0.23	-0.05	0.00	0.00	-0.23	-0.01	-0.01	0.00
Asian	-0.02	0.05	0.03	0.00	-0.06	-0.01	-0.01	0.00
Black	0.02	0.00	-0.02	0.00	-0.03	0.01	0.01	0.01
Hispanic	0.02	0.03	0.01	0.00	-0.01	-0.01	-0.01	0.00
American Indian	-0.03	0.01	0.01	0.00	-0.09	0.00	-0.01	0.00
Multi Racial	0.03	0.02	0.01	0.00	0.06	0.00	0.00	0.00
White	-0.03	-0.04	0.00	-0.01	0.05	-0.01	-0.01	-0.01
Free Lunch	0.04	-0.02	-0.04	0.02	0.13	0.01	0.01	0.01
Is LEP	-0.04	0.01	0.01	0.01	-0.07	0.01	0.00	0.01
Disabled	-0.25	-0.03	0.01	0.00	-0.25	-0.01	-0.02	0.00
Gifted	0.17	-0.03	-0.01	-0.01	0.12	0.00	0.00	0.00
Old for Grade	-0.26	0.01	-0.01	0.00	-0.21	0.00	-0.01	0.00
Old for Grade * Free Lunch	-0.21	-0.04	-0.02	0.00	-0.14	0.00	0.00	0.00
MS Mobility	-0.05	-0.05	-0.03	0.00	0.08	-0.03	-0.03	0.00
Achievement								
Passed Alg 1 in MS	0.08	0.10	0.09	-0.02	0.10	-0.03	-0.03	-0.02
Middle Sch. Avg Math Scr.	0.41	0.02	0.04	-0.02	0.28	-0.01	0.00	-0.01
Middle Sch. Avg Reading Sc.	0.44	0.00	0.05	-0.01	0.34	-0.01	0.00	-0.01
Grade 8 Science Score	0.33	0.04	0.09	-0.01	0.33	-0.01	-0.01	-0.01
Absences								
Middle Sch Avg Days Absent	-0.21	-0.03	-0.02	0.01	-0.18	0.00	-0.01	0.00

Notes: The entries in the table shows 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