Impacts of School Reforms in Washington, DC on Student Achievement

October 30, 2020

Dallas Dotter, Duncan Chaplin, and Maria Bartlett

Mathematica
1100 First Street, NE, 12th Floor
Washington, DC 20002-4221
Phone: (202) 484-9220
Fax: (202) 863-1763
Reference Number: 50580

Abstract

In 2007, the District of Columbia (DC) began a systemic reform of educational governance and processes that sought to produce dramatic improvements in student outcomes. These reforms included implementation of a more rigorous staff evaluation system, steady growth of the public charter school sector, and the introduction of a unified enrollment system. We estimate the cumulative impacts of these reforms by analyzing how changes in achievement levels of DC schools compare to changes observed for similar students in similar geographic areas without such reforms. Our analysis improves on prior efforts to study these reforms in several ways. We use nearly a quarter century of data (from the early 1990s to 2017), which enables us to cover more cohorts of students than previous studies—including achievement in grades 4 and 8 for five cohorts of DC students before 2007 and three cohorts after. We also take advantage of recent advances in constructing counterfactual outcomes in situations where one or very few units are treated. We find that the reforms in DC were associated with larger than expected growth in grade 4 math and reading scores on the National Assessment of Educational Progress. We also find similar gains in grade 8 math, especially for cohorts with more exposure to the reforms, but not in grade 8 reading. These results suggest that the reforms improved math education in kindergarten through grade 4 with impacts lasting to grade 8. At one-third of a standard deviation for math, the impacts we find in DC are similar in magnitude to those observed for math in New Orleans, where major school reforms were implemented starting in 2006–2007, immediately after hurricane Katrina, and larger than for some well-known education interventions like Success for All and the class size reductions in Tennessee. The results are less clear for reading and for achievement in high school, where data limitations precluded a credible impact analysis.

Acknowledgments

We thank the following individuals for their support and guidance on this project. These include many of our colleagues at Mathematica (Brian Gill, Josh Haimson, Phil Gleason, Steve Glazerman, Emma Ernst, Jennifer Brown, and Anuja Pandit), Stuart Buck at Arnold Ventures. Financial support for this work came from Arnold Ventures, which has supported initiatives designed to improve and expand educational options in several cities, including Washington, DC.

Contents

Tables

Figures

I. Introduction

In 2007, the District of Columbia (DC) began a process of school reform with the Public Education Reform Amendment Act (PERAA). Specifically, PERAA led to several important reforms, including changes in school governance structures, human capital policies, and leadership. Three key reforms in DC included (1) mayoral control of the DC Public Schools (DCPS) in 2007, which famously led to the implementation of high-stakes accountability for teachers through the IMPACT staff evaluation system in 2009; (2) legislation giving the DC Public Charter School Board sole authority over all public charter schools in DC, under which steady growth of the public charter school sector continued but with an increased focus on accountability for academic performance (NRC 2015); and (3) the introduction of a unified enrollment system for DCPS and public charter schools by 2014, which facilitated school choice by enabling parents to more fully take advantage of all of their DCPS and public charter school options. Many of these changes are in line with models for reforming school governance that are intended to increase the quality and diversity of schooling options through citywide offerings of more autonomous schools in the public charter and traditional public-school sectors.

Implementing high-stakes accountability for teachers might improve teacher quality by supporting improved performance of existing teachers and through dismissals of less effective teachers. The IMPACT evaluation system sought to increase the rigor of evaluation methods for teachers by incorporating value-added measures of their contributions to student achievement as well as in-depth classroom observations. These data were used for both teacher retention and promotion decisions.

Public charter schools might improve student outcomes both directly, by using innovative methods not available in other schools, and indirectly, by inducing positive competitive responses from other schools. DC experienced a steady increase in the fraction of students attending public charter schools, with roughly 46 percent of public-school students in DC enrolled at a public charter school by 2017, compared to 27 percent in 2007.

One way to promote competition between schools is to allow parents to choose a school among many options. DC has allowed within-district transfers since at least 1996 (DC 2006; Hurst et al. 2008). To enhance the benefits of school choice by making it easier to apply to many schools, some districts have implemented unified enrollment systems. DC implemented such a system starting in the 2014–2015 school year.

This study improves on past efforts to estimate the impacts of school reforms implemented in Washington, DC, after 2007. Specifically, we estimate (1) how test scores and student demographics in DC changed over time after 2007, compared to similar students in geographic areas without such reforms; (2) how results differed by student demographics; and (3) how postsecondary readiness among DC students changed in terms of SAT participation and achievement.

To estimate impacts of the reforms in DC, we address several challenges. First, DC did not implement a test that can be used to evaluate performance relative to other cities. This is both because DC is not part of a larger state containing other cities using the same assessment and because DC did not use the same assessments as any cities in other states around the time of the reforms of interest. To address this challenge, we use data from the National Assessment of Educational Progress (NAEP) that have been collected nationwide for decades. A second challenge is that NAEP data do not follow individual students. To address this, we estimate growth among pseudo cohorts based on differences in NAEP achievement between students sampled in grade 8 and students sampled in grade 4, four years earlier. A

third challenge is that we have only one geographic unit in our treatment group—Washington, DC. To overcome this, we construct counterfactual outcomes for DC using geographic areas without similar reforms and use these counterfactuals to estimate impacts on NAEP achievement. A fourth challenge is that NAEP provides insufficient historical data on grade 12 to assess the effects of reforms at the high-school level.[1] As an alternative, we explore using SAT data to capture student achievement after grade 8. A final challenge is that student demographics in DC have shifted over time, confounding estimated impacts of the reforms in DC. To address this, we use data on student demographics to control for and explore the possible role of changing student compositions on estimated impacts.

We find that grade 4 NAEP achievement in DC improved in both math and reading, relative to grade 4 achievement in otherwise similar settings without PERAA reforms, and that for math, these improvements increased with years of exposure to the reforms. However, similar improvements were not seen in grade 8 reading, suggesting that the improvements observed in grade 4 might have simply displaced learning that would have occurred in later grades in the absence of the reforms. In math, the improvements in grade 4 persisted in grade 8 but did not increase, suggesting that instructional improvements may have been primarily in the early grades. At the high-school level, we were unable to produce credible estimates of the impacts of reforms, because a large fraction of students did not take the SAT in most of the relevant years, which leaves more room for bias (see Appendix A, Section III for analyses of SAT participation and achievement).

The remaining sections of this paper are structured as follows. We summarize earlier literature looking at impacts of school reforms in Washington, DC, in Section 2. This is followed by a discussion of the data we use to analyze these reforms (Section 3). Our use of methods to estimate causal impacts when there is a single treated unit is covered in Section 4, followed by our results in Section 5. Section 6 concludes with a discussion about the implications of our results and avenues for future research.

[1] Grade 12 was not added to the state NAEP sample until 2009, and even then, only 11 states participated (National Center for Education Statistics 2010).

II. Evidence on School Reforms in Washington, DC

DC presents a unique set of challenges for studying school reform because it is both a city and, for purposes of education policy, a state. Attempts to estimate impacts of the reforms that started in 2007 in DC by comparing standardized testing outcomes in DC with those in nearby non-DC schools are hampered because the state assessment system in place during most of the relevant period, known as the DC Comprehensive Assessment System (DC-CAS), was administered only in DC.[2] Studies using DC-CAS scores as student outcomes have therefore either focused on describing trends over time (Education Consortium for Research and Evaluation 2014) or focused on impacts of mediating factors such as teacher turnover and teacher effectiveness (Adnot et al. 2017). Other studies have relied on variation between schools in the size and timing of the DC reforms' impacts on factors that mediate student achievement, looking at the impact of these factors on student outcomes as an indirect way of studying the reforms. For example, Walsh and Dotter (2014) use the spike in principal dismissals after the 2007 DC reforms to examine the impact of principal replacements on student achievement in DC, compared to the achievement of similar students whose principal was not dismissed.

Past efforts to understand the impacts of school reforms on student outcomes use a variety of methods and outcomes, generally with limited use of controls to distinguish the true impacts of reforms from other factors whose correlation with school reform might be coincidental or spurious. Although some of these studies focus on a single city (e.g., Harris and Larsen 2019), many have looked at statewide reforms, comparing the outcomes of students in states that implemented reforms with outcomes in other states that had not yet enacted such reforms. Because such comparisons require a common assessment across states, these studies tend to use results from the NAEP, which tests random samples of U.S. students in traditional public, public charter, and private schools at grades 4, 8, and 12 every few years.

To estimate the overall impacts of reforms in DC, a researcher must have a comparison group from outside the district with comparable measures. NAEP data can be used for this purpose. Several studies use NAEP and other test score data to look at student achievement in DC over time, with mixed degrees of credibility, often without a comparison group, and always using far fewer years of data than we cover either before or after PERAA was implemented in 2007.

A few of these studies provide suggestive evidence of limited impacts from the reforms that started in 2007. For example, a 2011 National Research Council (NRC) evaluation focuses on trends in average NAEP scores from 2002 through 2009 in DCPS and in several other urban school districts. This descriptive analysis shows that from 2007 to 2009, DC experienced increases from one cohort to the next in NAEP scores for grade 4 math and reading and for grade 8 math, which is sometimes interpreted to mean that the reforms worked. However, those same analyses also show that these increases in DC started around 2003 and that similar increases occurred outside of DC. In a similar vein, Weiss and Long (2013) use data from 2003 to 2011 and claim that DC reforms failed to improve student outcomes, noting the lack of a rise in the nationwide rankings of DC's NAEP scores immediately after reform implementation. However, their data do not include NAEP scores after 2011.

Other studies have found evidence suggesting that the reforms did matter. For instance, Özek (2014) and Blagg and Chingos (2016) find that NAEP scores in DC rose by more than would be expected based on the demographic shifts between about 2005 and 2013. Özek (2014) finds a similar pattern in the state

[2] The Partnership for Assessment of Readiness for College and Careers (PARCC) exam was first administered in DC in spring 2015—too late to be used in our study design.

assessments used in Washington, DC. However, neither paper compares achievement in DC with geographic areas elsewhere. Carnoy, Garcia, and Khavenson (2015) find that DC scores rose relative to other states between 1992 and 2013, but they do not use a matched comparison group or rigorously estimated counterfactual and do not focus on comparisons before and after PERAA was implemented. Osborne and Langhorne (2018) focus on the post-PERAA period and show that DC improved NAEP achievement compared to other large cities between 2007 and 2017, although their work does not adjust for trends in achievement prior to PERAA.

Our work builds on the studies above by using rigorous methods to estimate counterfactual outcomes for DC and by using more years of data. Our methods are designed specifically to account for pre-policy trends both in DC and elsewhere. Equally importantly, by starting earlier than most of the studies above and continuing for much longer, our data can better distinguish impacts of PERAA from pre-existing trends and determine whether impacts might materialize in the longer term, even if they do not seem to show up immediately after PERAA went into effect.

III. Data Sources and Sample Definitions

In this section, we first describe the two types of panel data we use to estimate our models and then describe the NAEP test score data that we use for student outcomes. We create panel data by using geographic areas within which we can follow NAEP scores over time. We use two different units of analyses, estimating our effects of interest separately for each: state-level NAEP performance over time and county-level NAEP performance over time. Although the state-level data provide far more precision per unit than the county-level data, the county-level data enable us to use within-state variation, potentially resulting in counterfactuals that better reflect DC.

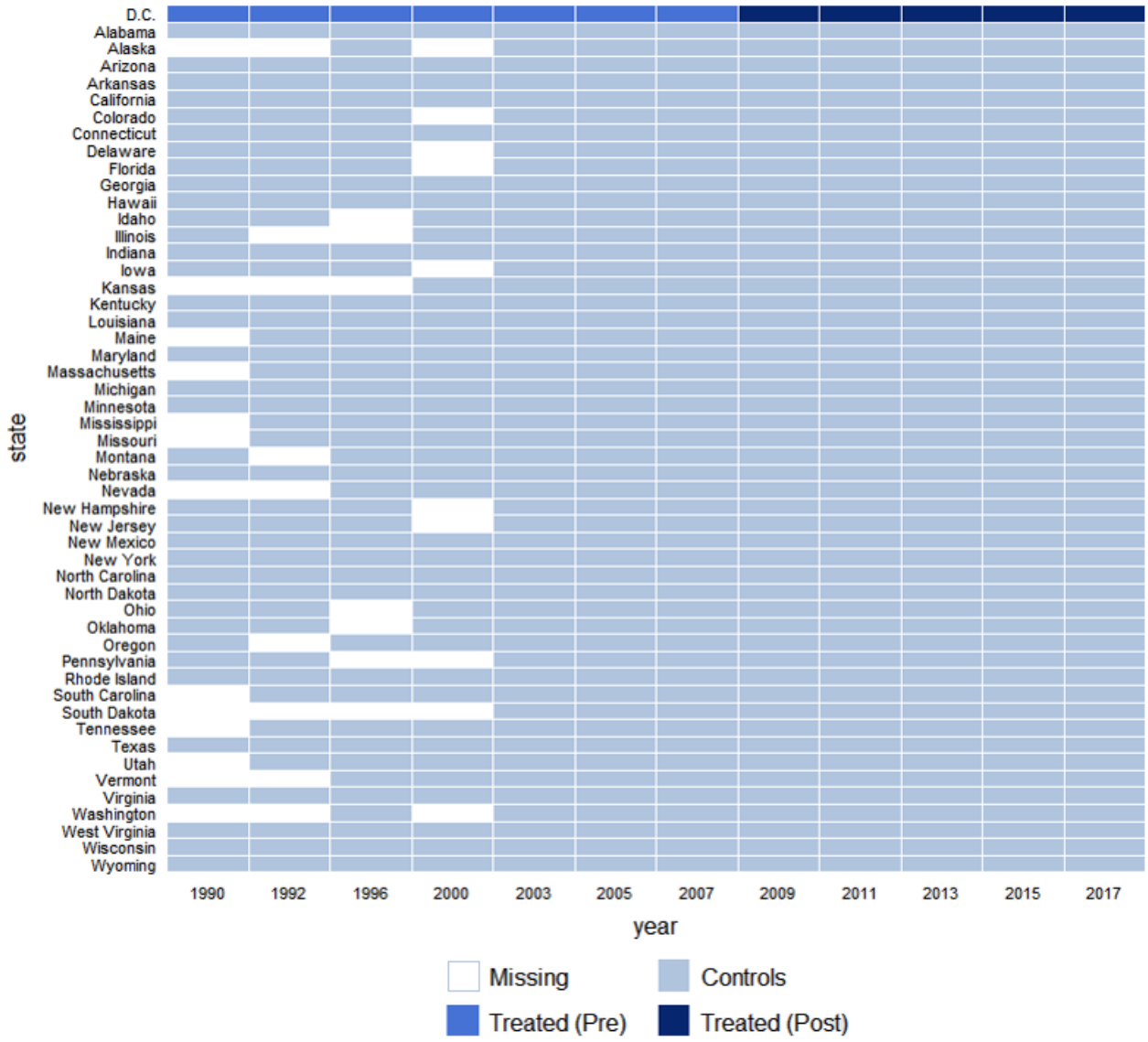
A. Panel data on NAEP scores, by state and county

State-level data. We use state-level data publicly available from the National Center for Education Statistics (NCES) online as well as data from the Urban Institute's interactive NAEP dashboard.[3] The latter are constructed from restricted-use student-level NAEP scores, with and without adjustments for differences in the characteristics of students sampled by NAEP across states and years (Chingos et al. 2019).[4] Because of variation in states' NAEP participation prior to 2003, constructing panel data of state NAEP scores without missing values involves striking a balance between the number of comparison states and the number of years included. For example, given the pattern of state participation in grade 8 NAEP for math shown in Figure III.1, such choices include 12 cohorts and 24 states, 11 cohorts and 31 states, or 10 cohorts and 35 states (not including DC). We selected 10-cohort panels because counterfactual estimates for untreated states had the lowest average mean squared prediction error across grade 8 math and reading using this panel (see Section 4 for details on prediction error and Appendix B Table B.1 for prediction errors across subjects and panel sizes). These panels cover the years 1996 through 2017 for math and 1998 through 2017 for reading. The exact number of states contributing to the estimated counterfactual in a given model depends on the number of states with available data in the 10-cohort panel for that grade, subject, and student subgroup when applicable.

County-level data. We create a panel of county-level data using NAEP restricted-use student-level data aggregated to county-level units. We use counties rather than districts to explore within-state variation because the restricted-use NAEP data do not consistently capture district identification in many early years. We considered using the Trial Urban District Assessment (TUDA) NAEP data. Although that collection now covers many urban districts, TUDA did not start until 2002, covered few districts in its early years, and perhaps most importantly, has not covered the many public charter schools within DC (NCES 2018b). For grade 4 math, using restricted-use NAEP data we are able to include about 60 counties, each of which had at least eight schools sampled every year from 1996.[5] We obtain about the same number of counties for grade 4 reading, about 40 counties in grade 8 math, and about 50 in grade 8 reading.

[3] The NCES NAEP data explorer can be accessed at https://www.nationsreportcard.gov/ndecore/xplore/nde.
[4] See http://apps.urban.org/features/naep/ for the Urban Institute's "America's Gradebook" NAEP dashboard. The corresponding technical appendix describing the data is available at http://apps.urban.org/features/naep/naep-technical-appendix.pdf.
[5] These sample sizes are based on the restricted-use data we analyzed for grade 4 math. Although NAEP restricted-use data do not always provide county information, we are able to recover missing county identifiers from the corresponding years of the Common Core of Data (CCD) using CCD school identification numbers (IDs), which are provided in NAEP in the years when the county identifiers were missing.

Figure III.1. Grade 8 NAEP math score data availability by state and year



Source: State-level grade 8 NAEP scores from 1990 through 2017.

NAEP = National Assessment of Educational Progress.

B. NAEP sample design and coverage

The NAEP data we use include many years and schools. The data span a quarter of a century, from 1992, 15 years before PERAA was implemented, to 2017, a decade after PERAA. We focus on public schools in the NAEP data, including public charter schools. The data cover at least 100 schools in DC each year for grade 4 and at least 30 schools for grade 8.[6] Outside of DC the NAEP data we use for our state panel

6 These numbers are based on publicly available data such as https://files.eric.ed.gov/fulltext/ED369067.pdf and https://nces.ed.gov/nationsreportcard/tdw/sample_design/2011/2011_sampdsgn_state_schlresp_gr8.aspx for 1992

cover at least 3,000 grade 4 and 3,000 grade 8 schools each year.[7] For our county panel, we restrict our analytic samples to schools in counties from which at least eight schools were sampled in each year of the NAEP data we use.[8] We create the county panel by aggregating these data to the county level. The resulting sample sizes outside of DC in our county panel remain large, ranging from roughly 1,000 to 2,300 schools per year for grade 4 and roughly 630 to 1,400 schools per year for grade 8.[9] Roughly 60 students per sampled school were chosen to participate in the NAEP assessment across math and reading (NCES 2017).

As shown in Figure III.2, restricted-use NAEP data for math are available at the student level every two years from 2003 to 2017 and every four years from 1992 to 2000, meaning that 10 cohorts of grade 8 students can be followed from grades 4 to 8. We have five cohorts of students who were in grades 4 to 8 before the reforms began in 2007 (cohorts 1 through 5) and three cohorts who were in those grades after that time (cohorts 8, 9, and 10). The remaining two cohorts (6 and 7) attended grades 4 through 8 just as the reforms were implemented. Hence, we describe the relationship between impacts for these cohorts and the length of time for which they overlap with reforms that were implemented. A similar set of cohorts is available for NAEP scores in reading (see Figure B.1 in Appendix B).

and 2011, respectively. We provide numbers of counties in our county-level panel in Appendix B, Table B.3. Numbers of states are provided in Table B.2.

[7] These sample sizes are based on publicly available data. The actual numbers of schools in NAEP are far greater in most years used in our analyses. However, the total number of schools in our sample is lower than the total available because some states did not have data available for all years included in our data, as shown in Figure 1.

[8] Another reason to use counties is that schools in a county with eight or more schools per year but in a district with less than 8 schools are included when using county units, but excluded when using districts.

[9] These sample sizes are based on the restricted-use data we analyzed. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1992-2017. Sample sizes are rounded to the nearest 10 to protect confidentiality.

Figure III.2. NAEP state-level math data: 10 cohorts of students in grades 4 through 8

	Grade levels by year and NAEP cohort										
Cohort	’92	’96	’00	’03	’05	’07	’09	’11	’13	’15	’17
1	4	8									
2		4	8								
3			4*	8							
4			4*		8						
5				4		8					
6					4		8				
7						4		8			
8							4		8		
9								4		8	
10									4		8

Note: Bold vertical line indicates implementation of reforms in 2007.

* Grade 4 scores in 2000 used for both 2003 and 2005 grade 8 cohorts, representing three- and five-year gaps, respectively.

NAEP = National Assessment of Educational Progress.

For most of its history, NAEP has sampled large geographic areas (cities, counties, or groups of counties), schools within those areas, and students within schools. Stratification is used to ensure that certain groups, such as Washington, DC, are sampled with certainty. We use the weights provided by NCES to adjust for oversampling and bootstrap estimates to adjust for uncertainty caused by the sampling. We use bootstrapping instead of the replicate weights provided by NCES because the replicate weights correspond to sampling units that differ by year; therefore, they do not align with our sample of counties containing at least eight schools in the NAEP data each year. Also, since we use aggregate data at the state or county level, our methods automatically adjust for clustering of students within these units, thus accomplishing one of the major benefits of using the replicate weights.[10]

Our analyses also adjust for the complicated nature of how NAEP is implemented in the classroom. To reduce the amount of time required for testing, each student answers only a subset of the questions on the NAEP tests. The results for each student are combined with results for other students and used to create several “plausible values” for each student. We describe how these plausible values are used for our estimates in the empirical strategy section below.

C. Identifying areas with reforms similar to those in DC

We estimate impacts using models that include all possible comparison units, including areas that implemented similar reforms as well as models using a restricted sample that excludes such units. We use the restricted sample in order to estimate impacts using the difference between observed outcomes for DC

[10] Replicate weights also account for stratification and units that were sampled with certainty. This does not matter for us because we are implicitly treating all units as if they were sampled since we are only observing a subset of all possible outcomes—treated outcomes for the treatment group, and untreated outcomes for the comparison group.

in the presence of the PERAA reforms and estimated outcomes for DC in the counterfactual case where the PERAA reforms were not implemented. We exclude from the restricted sample units (states or counties) where similar reforms were implemented during the same timeframe. For example, we exclude the county containing New Orleans, where similar reforms were undertaken starting in 2005 in the wake of hurricane Katrina (Harris and Larsen 2019).

Measures used to identify restricted sample. Our selection of units to exclude from the restricted sample is based on the implementation of reforms between 2007 and 2014. If a unit's district(s) implemented some of these reforms before 2007 and then stopped, we include them in the restricted sample. Also, if a unit had districts that started to implement these reforms after 2014, we include them in the restricted sample based on the theory that they implemented the reforms too late to influence outcomes during the years included in our sample. We excluded units if most students in that unit were in districts implementing one or more of the relevant reforms. We ultimately drop four states from our state-level analyses based on teacher tenure reforms that had been implemented. We drop very few counties beyond those four states based on any of the reform areas. To maintain confidentiality of NAEP participating schools, we do not name which counties are dropped in our discussion below. Rather, we only discuss which counties would have been dropped, had they been in the NAEP data used in our full models.

Teacher tenure. The reforms in DC had the effect of reducing the benefits of tenure by making it easier to fire tenured teachers based on poor performance (Gitomer et al. 2014). We exclude four states with laws designed to end teacher tenure between 2007 and 2014 from our state-level models and all counties in those states from our county-level models. We identify these states using state- and district-level data provided to us by the National Council on Teacher Quality (NCTQ) combined with web searches to verify those data and the specific years when the no tenure policies were in place. The exact definition of teacher tenure is ambiguous, but the wording used in the NCTQ data was clear (Christie et al. 2010). We search the NCTQ records for the phrases "no tenure," "abolishment of tenure," "tenure is non-existent," "tenure non-existent," "tenure does not exist," and "only awards annual contracts." We identify eight states as having no tenure at some point in time based on this review. Further review of the NCTQ district records and web searches suggests four of these states (Florida, Kansas, North Carolina, and North Dakota) as having a policy of no teacher tenure during our period of interest. Therefore, we exclude these four states from our restricted sample.

Public charter enrollment. We exclude one county and no states based on the rate of public charter enrollments. In order to determine which counties to omit based on this characteristic, we analyze enrollment by school type and district using the Common Core of Data. As noted earlier, public charter school enrollment in DC increased from 27 percent in 2007 to 44 percent by 2014. New Orleans also experienced a large growth rate, going from close to zero percent in 2004, before hurricane Katrina (Jacobs 2015), to more than 90 percent by 2014 (National Alliance for Public Charter Schools 2014). No other county with data on 10 or more consecutive NAEP cohorts experienced a similarly large increase.[11] Thus, we ultimately omit only New Orleans from inclusion in the restricted sample based on public charter school enrollment.[12]

[11] Some cities with larger public charter enrollment rate growth (for example, Detroit, Flint, and Kansas City) were parts of larger counties that did not experience such growth countywide. Other cities with large public charter growth, like Philadelphia, were excluded because the entire state was missing from some years of the NAEP data.
[12] New Orleans would have also been omitted from the restricted sample because of its use of unified enrollment.

Unified enrollment. We exclude three counties and no states based on the use of unified enrollment. As noted earlier, DC implemented a unified enrollment system starting in the 2014–2015 school year. Hesla (2018) identified three other cities (Denver, New Orleans, and Newark) that had similar systems in place by 2014–2015, so the counties containing those cities were excluded from our restricted sample.

Measures not used to identify sample restrictions. We considered two other measures that we ultimately do not use when identifying the restricted sample.

Principal accountability. The IMPACT system changed principal accountability in DC. However, the No-Child-Left-Behind (NCLB) legislation of 2002 mandated strong accountability at the school-level that may have also had important incentive effects on principals. Indeed, most states sanctioned schools based on their performance, even before NCLB (Hurst et al. 2008). Thus, DC was not unique in terms of principal accountability.

Teacher performance measures and use. We also explored restricting our sample based on the use of value-added measures to rate teachers or otherwise hold them accountable for student achievement. However, we found that the NCTQ data have far too many ambiguous results on this topic, perhaps due to difficulty NCTQ staff may have faced when trying to sort through district records. A quick summary of the NCTQ data suggests that half of the districts we might have included in our restricted sample may have used either value-added or some other growth measure to evaluate teachers because they answered positively to a question about using either growth or achievement for teachers of tested and untested subjects (assuming any response to questions mentioning test scores without also mentioning goals or targets could refer to the use of a growth measure). NCTQ also includes a question about the relationship between evaluation ratings and annual salary increases. Of 230 records on that question, 121 state, “Issue not addressed....” All other records suggest that there is, or there could be, a relationship between evaluation ratings and teacher salary. Hence, in the NCTQ data DC does not stand out compared to other districts based on how much teacher salary varies with performance.

IV. Empirical Strategy

Our empirical strategy estimates the effects of reforms implemented in DC as the difference between outcomes for NAEP cohorts observed in DC post-implementation and the expected but unobserved outcomes for those cohorts in DC in the same year(s), but in the absence of the reforms that were implemented. Our primary outcomes include NAEP achievement in math and reading for each cohort of grade 8 and grade 4 students sampled by NAEP, as well as within-cohort gains in NAEP achievement between grades 4 and 8. We define the average effect of those reforms in year *t* as

$$(1) \qquad \tau_{DC,t} = Y^1_{DC,t} - Y^0_{DC,t} \ ,$$

where Y^1 and Y^0 represent outcomes Y under treated and untreated states, respectively. Because one cannot observe outcome values for DC in post-PERAA years without the influence of the reforms implemented, we focus on estimating $Y^0_{DC,t}$ for all periods *t* after 2007. These estimated counterfactual outcomes are then used to calculate treatment effects as

$$(2) \qquad \hat{\tau}_{DC,t} = Y^1_{DC,t} - \hat{Y}^0_{DC,t} \ ,$$

where $\hat{Y}^0_{DC,t}$ is an estimate for the counterfactual outcome in DC during period *t*. An overall average treatment effect is calculated as the average of estimates $\hat{\tau}_{DC,t}$ across all *t* after 2007.

A. Estimation of counterfactual outcomes

We estimate counterfactual outcomes for DC using a method for estimating causal effects referred to as matrix completion (MC). MC and related counterfactual estimation methods have received recent attention in a body of literature generalizing different approaches to estimating causal effects using panel data. These methods nest approaches such as difference-in-differences and synthetic controls as special cases (Bai 2009; Doudchennko and Imbens 2016; Gobillon and Magnac 2016; Xu 2017; Athey et al. 2018; Arkhangelsky et al. 2019). MC considers unobserved counterfactual outcome(s) to be missing elements in a matrix of outcomes for all *N* units and *T* periods in the absence of treatment. Specifically, MC estimates the "missing" unobserved counterfactual outcomes by using all observed data in that matrix (i.e., all observations except those for the treatment group in the treatment periods) to solve the minimization problem

$$(3) \qquad \min_{\boldsymbol{M},\mu,\gamma} \sum \left(Y_{it} - \mu_i - \gamma_t - M_{it}\right)^2 + \lambda \left\| M \right\|_* ,$$

for all (*i*,*t*) that are not the treated unit in the treatment periods. Here, Y_{it} is the outcome of interest for unit *i* at time *t*; μ_i is a fixed effect for unit *i*; γ_t is a common effect across units for time *t*; M_{it} represents both observed and unobserved time-varying factors (for example, unobserved time-varying confounders); and $\lambda \left\| M \right\|_*$ is the penalty term that imposes a cost on model complexity to the minimization. $\boldsymbol{M}$ can be represented as a factor model that captures unit-specific, time-varying components of the underlying data-generating process, after accounting for unit fixed effects μ_i and common period effects γ_t, using

variation in the data to identify how many factors appear to be present (the rank), their unit-specific coefficients (loadings), and values of the time-specific factors. This factor model is of the form

$$(4) \qquad M_{it} = \sum_{r=1}^{R} L_{ir} F_{tr}$$

where L_{ir} are unit *i*'s separate loadings for each of the *R* factors specific to year *t* represented by F_{tr}. MC uses matrix factorization—specifically, Singular Value Decomposition (SVD)—and a nuclear norm regularization to estimate the rank *R* and matrix *M* in equation (4).[13] We use cross-validation to find the optimal penalty weight λ (following the approach of Athey et al. 2018). Counterfactual outcomes for the treatment group are estimated by using predicted values from the equation,

$$(5) \qquad \hat{Y}^{0}_{i,t} = \hat{\mu}_i + \hat{\gamma}_t + \hat{M}_{it}$$

for all (*i*,*t*) for the treated unit in the treatment periods. This generalized approach has several attractive qualities. First, like synthetic control methods, it can accommodate designs such as ours, in which there is one treatment unit but many potential comparison units. Second, compared to traditional synthetic controls such as those in Abadie, Diamond, and Hainmueller (2010), this method more flexibly incorporates variation in outcomes between units as well as variation within treated units over time to estimate counterfactual outcomes. The resulting counterfactual outcome values often better match those of the treatment group in the pre-treatment period compared to common difference-in-differences approaches, which often fail to satisfy their underlying parallel trends assumption when there is one or few treated units.[14] Third, whereas difference-in-differences requires many units relative to time periods and traditional synthetic controls require many time periods relative to units, MC uses regularization and is flexible in the dimensions of panel data that can be used. Fourth, by estimating the number of factors in the model, this method relaxes the rigid functional form imposed by a standard difference-in-differences model and can better approximate more complex underlying data-generating processes.

B. Estimation uncertainty

Our analyses estimate impacts on a single treated unit (Washington, DC). Doing so prohibits the calculation of traditional standard errors that require information on multiple treated units. To describe the level of uncertainty around our estimates, we follow the literature in using three estimates of uncertainty commonly used for synthetic control methods—(1) root mean squared prediction error (RMSPE) statistics based on placebo treatment effects for untreated units, (2) 95 percent confidence intervals, and (3) fit-adjusted *p*-values that adjust for model fit in pre-treatment years. In each case we are implicitly assuming that the variance in counterfactual outcomes for the treatment group can be approximated by using the variance in outcomes observed among the comparison group.

RMSPE. We estimate the uncertainty with which the methods used can estimate counterfactual outcomes in the post-treatment period by performing placebo tests using untreated units from the control group and summarize this information using an RMSPE statistic. We remove DC from the sample and then,

[13] See Hastie et al. (2015) for a discussion on matrix completion algorithms using singular value decomposition and nuclear norm regularization (also referred to as a "soft impute"), and Athey et al. (2018) for extensions relating to panel data applications for estimating treatment effects.
[14] See Arkhangelsky et al. (2019) for a discussion of difference-in-differences parallel trend assumption versus synthetic control-type weighting approaches.

separately for each unit, assume that unit was treated after 2007 and repeat the process used to estimate counterfactual outcomes for DC. Because we do observe the untreated outcomes for each of these units, the difference between their estimated and observed values provides a measure of how well the method can predict outcomes in the absence of treatment. For each outcome, we use this difference across all units' placebo tests for all years after 2007 to calculate the RMSPE for that outcome as

$$\sqrt{\frac{1}{N_{i\neq DC}\left(T-T0\right)}\sum_{i\neq DC}\sum_{t=T0+1}^{T}\left(Y_{i,t}^{0}-\hat{Y}_{i,t}^{0}\right)^{2}} \quad (6)$$

where *T0* is the last pre-treatment year and *T* is the final year of panel data. Note that if the difference between outcomes predicted by placebo tests and actual outcome values observed among untreated units is zero in expectation, the RMSPE statistic is analogous to the standard error of counterfactual estimates among untreated units. To the extent that our counterfactual estimator performs as well for DC as it does on average for other states, it provides a reasonable estimation of a standard error for the predicted counterfactual outcome for the treated unit (as well as for the treatment effect, after re-centering). We report the RMSPE statistics in all our tables.

To reduce the amount of time required for NAEP testing, each student answers only a subset of the questions and each students results are combined with other students' results to create several "plausible values" per student.[15] We use these plausible values to adjust all estimates based on the county-level data. More precisely, we estimate each model separately for each of the five plausible values for student's scores in a given subject and grade level. We combine the resulting five estimates using their mean for point estimate and RMSPE values, and use the multiple imputation method described by Rubin (1987) for the variance of estimates, which combines the estimate's within-plausible value variance and the variance of estimates across plausible values.

Confidence intervals. We use year-specific RMSPE statistics to construct 95 percent confidence intervals, assuming normality, around post-treatment counterfactual estimates for all figures. In practice these RMSPE-based confidence intervals are considerably wider than confidence intervals based solely on the distribution of bootstrapped estimates. However, we bootstrap both counterfactual estimates and their corresponding RMSPE statistics to reduce potential bias contained in any single estimate for an untreated unit. In the county-level data these confidence intervals account for the complicated sampling in the NAEP data associated with the facts that each student takes only a subset of the test items and that students are sampled using a multistage sampling design, as discussed earlier.

To implement the bootstrapping procedures we repeatedly draw random samples of untreated units with replacement from the original sample, maintaining the original sample size. We produce counterfactual estimates and corresponding RMSPE using each of these samples, forming distributions that reflect the variance of each that is attributed to the sampling of untreated units used for estimation. We use 500 bootstrap iterations for results using state-level units and 100 bootstrap iterations for each of the five plausible outcome values when using county-level units, also resulting in 500 iterations per estimate.

Fit adjusted *p*-values. Finally, we use the empirical distribution of treatment effect estimates among DC and untreated units (via placebo tests), adjusted for the pre-treatment fit, to approximate the likelihood of

[15] NCES provides 5 plausible NAEP score values for years before 2011 and 20 plausible values for years 2013 and later. For those later years, we select and use 5 of the 20 plausible values to maintain a consistent process for computing point estimates and variance across years. Specifically, we choose the 1st, 5th, 10th, 15th, and 20th plausible value using the ordered labels provided by NCES.

observing a particular unit's estimate by random chance. Specifically, we use the distribution of a statistic commonly used for synthetic control approaches that (1) captures the magnitude of differences between a unit's observed outcomes and estimated counterfactuals in the post-treatment period and (2) down-weights that magnitude in inverse proportion to similar difference in the pre-treatment period, thus penalizing post-treatment estimates when the model poorly fits the observed untreated outcomes in earlier periods.[16] For DC, we calculate

$$\phi_{DC} = \frac{\sqrt{\frac{1}{T-T0}\sum_{t=T0+1}^{T}\left(Y_{DC(1),t} - \hat{Y}_{DC(0),t}\right)^2}}{\sqrt{\frac{1}{T0}\sum_{t=1}^{T0}\left(Y_{DC(1),t} - \hat{Y}_{DC(0),t}\right)^2 + 1}} \quad (7)$$

whereas for an untreated unit *i*, ϕ_i is calculated similarly but uses the observed untreated values $Y_{i,t}^0$ in place of $Y_{DC(1),t}$. We use the empirical cumulative distribution of ϕ_i, $F\left(\phi_i\right)$, to derive the *p*-value for a two-tailed test of the null hypothesis of no treatment effect for a particular unit *j*:

$$\hat{p}_j = 1 - F\left(\phi_j\right);\ F\left(\phi_j\right) = \frac{1}{N}\Sigma_i 1\left\{\phi_i < \phi_j\right\} \quad (8)$$

[16] See, for example, Abadie et al. (2015). Our method is similar to what has been done in the literature, except that we add 1 to the denominator to avoid exceptionally large statistics driven almost entirely by small deviations from near-perfect fits in the pre-treatment period. With this adjustment, ϕ_i in equation (7) converges to the RMSPE for the post-treatment period as the error in the pre-treatment fit approaches zero.

V. Results

A. NAEP achievement

Our results suggest positive impacts on math scores in grade 4 that remain by grade 8.[17] Average treatment effects are roughly between 9 and 11 NAEP scale score points when adjusting for student characteristics (Table V.1). These translate to between one-quarter and one-third of a standard deviation (SD).[18] As shown in Figure V.1, NAEP scores in DC generally increased by more than the counterfactual year over year after 2007.

We also find positive impacts on NAEP reading scores for grade 4, but these are not sustained in grade 8. As shown in Figure V.1 impacts on grade 8 reading scores are only just above the 95 percent interval for the counterfactual estimate in 2015 but not in the other years. One interpretation of these results is that much of the gains in grade 4 reading were in skills that students might have picked up later in the absence of these reforms. Hence, those gains are not seen in grade 8 reading. This is supported by an analysis of within-cohort gains presented in Appendix Table A.2. Another possible explanation is that the reforms primarily affected achievement at the elementary level and the estimated grade 8 math impacts reflect residual benefits of earlier impacts, whereas the alignment between the grade 8 and 4 tests in reading is weaker than in math. However, the fact that the tests were originally designed to be comparable across grades suggests that this may not explain the lack of impacts for grade 8 reading (Camilli et al. 1993). A third possible explanation is that the grade 8 reading tests had too much bottom coding which ended up obscuring growth for DC students. This could be an area for possible future research.

Because there is very little difference between our estimates using states, counties, or the restricted sample versions of either, this section focuses on estimates using states without restrictions, given that they are slightly more precise than the other sets of results. A comparison of estimates using the different samples is presented in Appendix Tables A.1–A.4.

[17] We present results using NAEP scaled scores because those scores were originally designed for comparisons across grades (Camilli et al. 1993). See Appendix A for a discussion of scaling issues.
[18] The standard deviations of NAEP scaled scores in 2017 were 31 for grade 4 math, 39 for grade 8 math, 38 for grade 4 reading, and 36 for grade 8 reading (NCES 2018a, 2019).

Table V.1. Estimated impacts across years on NAEP achievement, by grade and subject

	Grade 8 Math	Grade 4 Math	Grade 8 Reading	Grade 4 Reading
Impact (NAEP scaled score)	9.4*	10.6*	1.0	8.8*
RMSPE	3.2	3.0	3.0	3.0
Fit-adjusted *p*-value	0.03	0.03	0.68	0.03
Effect size	0.30	0.36	0.09	0.28
N (states)	35	35	36	38

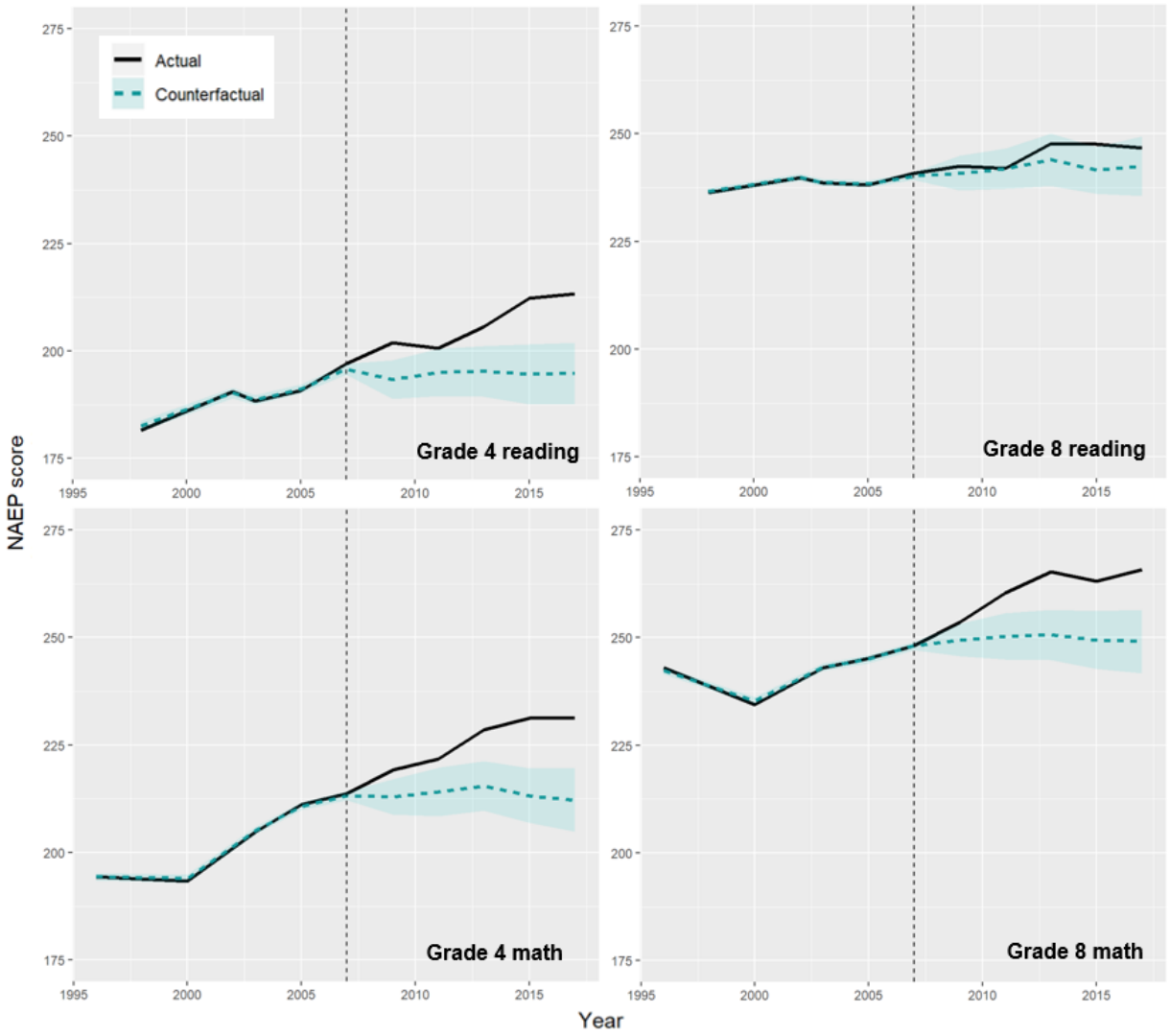
Source: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data, regression adjusted for student demographics.

Notes: Impacts and RMSPE units are in NAEP scaled score points. Fit-adjusted *p*-values are based on the samplewide distribution of the statistic described in our empirical strategy section, which is closely related to RMSPE and also model fit in the pre-treatment years.

* *p*-value < 0.05.

NAEP = National Assessment of Educational Progress; RMSPE = root mean squared prediction error.

Figure V.1. DC NAEP scores and estimated counterfactuals, by grade and subject



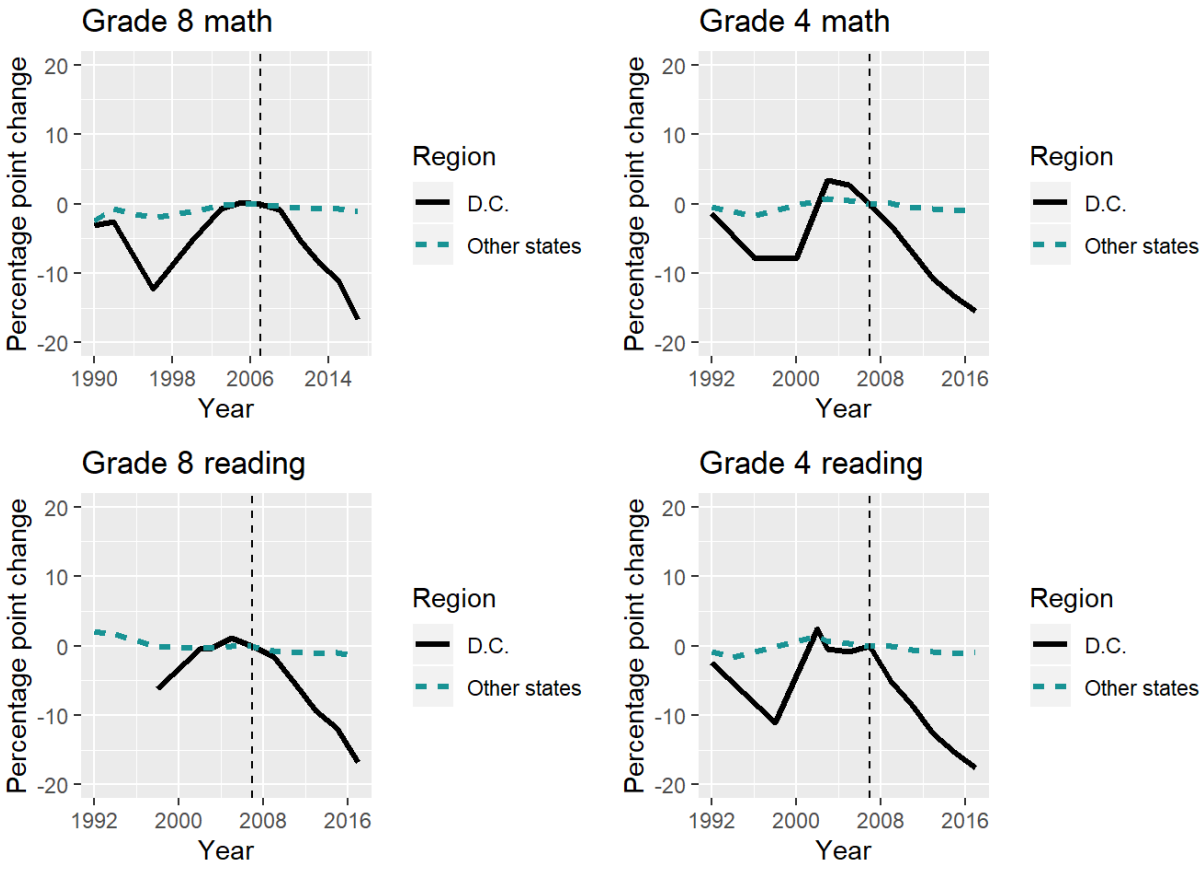
Notes: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data, regression adjusted for student demographics. Actual = observed NAEP scores for DC (solid line). Counterfactual = matrix completion estimates of counterfactual outcomes in the absence of reforms (dashed horizontal line). Vertical dotted line demarcates the implementation of PERAA in 2007. Shaded region is the 95 percent confidence interval based on the post-2007 RMSPE across all other states.

B. Demographic shifts in DC

Improvements in NAEP achievement for DC might be partially driven by the gentrification that occurred in DC over the period of reforms. Gentrification can directly affect achievement levels through changes in the composition of students in DC, and indirectly through peer effects as the composition of students' peers change. Indeed, the percentages of students in DC who are black and white have changed substantially over time and are reflected in the composition of students sampled by NAEP each year (Figures V.2a and V.2b). In particular, the percentage of students who are white has risen, whereas the percentage who are black has fallen, both by about 15 percentage points since 2007, relative to their percentages nationally. In contrast, there have been no clear changes, relative to national averages, in the

percentages of DC students who are Hispanic or of other races and ethnicities (see Figures A.1 and A.2 in Appendix A).

Figure V.2a. Percentage of black students in NAEP samples for DC versus other states, relative to 2007

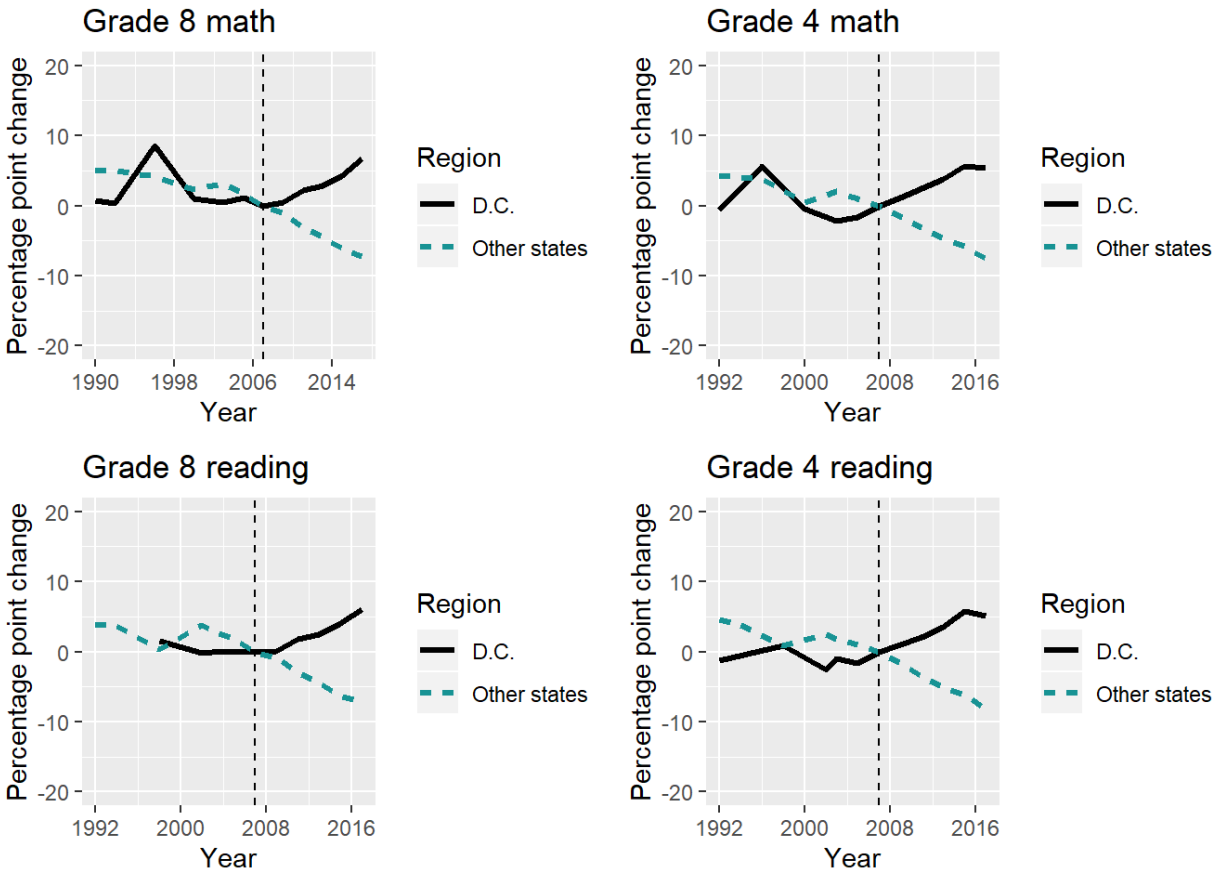


Source: Authors' estimates using state-level NAEP data.

Notes: These data cover the entire United States and not just the samples of counties and states used in our analyses. All percentages are set to the original values minus the value in 2007.

NAEP = National Assessment of Educational Progress.

Figure V.2b. Percentage of white students in NAEP samples for DC versus other states, relative to 2007



Source: Authors' estimates using state-level NAEP data.

Notes: These data cover the entire United States and not just the samples of counties and states used in our analyses. All percentages are set to the original values minus the value in 2007.

NAEP = National Assessment of Educational Progress.

To understand the extent to which DC's improvements in NAEP achievement were driven by gentrification versus improvements in education, we compare impact estimates that do and do not account for student demographics. We also separately estimate impacts for subgroups based on race and ethnicity. We do not control for free and reduced-price meal status in any of our runs given that (1) ways of coding that variable have changed over time and (2) a growing number of schools treat all students as eligible (Hewins et al. 2017). This is reflected in the data as a general upward trend over time for the percentage of students eligible for free or reduced-price school meals, in DC and the rest of the nation (Figure A.3 in Appendix A).

The overall pattern of the results in Table V.1 holds when using separate models by race (Tables V.2 and V.3). The estimated impacts are consistently smaller in magnitude than those that do not account for student characteristics. For example, the impact on grade 8 math in column 1 of Table V.2 is roughly 12 points overall and about 8 points among black students in column 3.

These comparisons suggest that compositional shifts of students in DC since 2007 were associated with average NAEP score increases of 3 to 4 points across subjects and grade levels. This corresponds to 0.09 to 0.12 SD and is consistent with expectations based on documented black–white NAEP achievement gaps of 0.6 to 0.8 SD (Reardon et al. 2014; Bohrnstedt et al. 2015) and the roughly 15 percentage point decrease in the proportion of DC students sampled for NAEP who were black (Figure V.2a). The smaller subgroup impacts are therefore plausible estimates of the impacts of the reforms, separate from the direct effects of changes in the composition of students after 2007.

Table V.2. Estimated impacts on NAEP math scores, by student race

	All Students		Subgroups		
	No controls	Demographic controls	Black	Hispanic	White
Grade 8 Math					
Impact (NAEP scaled score)	11.8*	9.4*	8.3*	9.5	n.a.
RMSPE	3.0	3.2	4.1	4.4	
Fit-adjusted *p*-value	0.03	0.03	0.04	0.07	
Effect size	0.30	0.24	0.21	0.24	
N (control states)	35	35	24	13	
Grade 4 Math					
Impact (NAEP scaled score)	12.9*	10.6*	9.5*	8.0	6.3
RMSPE	3.0	3.0	4.3	4.4	3.2
Fit-adjusted *p*-value	0.03	0.03	0.04	0.06	0.08
Effect size	0.42	0.34	0.31	0.26	0.20
N (control states)	35	35	27	17	36

Source: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data, regression adjusted for student race and ethnicity in column 2.

Notes: RMSPE units are NAEP scaled score points. Subgroup math scores for grade 8 white students in DC were not available via NCES because reporting standards were not met for this group in 1996, 2003, 2007, or 2009.

* p-value < 0.05.

n.a. = not available; NAEP = National Assessment of Educational Progress; RMSPE = root mean squared prediction error.

Although we cannot directly account for peer effects as a result of changing student compositions, the literature suggests these effects are relatively small. Hanushek, Kain, and Rivkin (2009) find achievement among black students in Texas was on average 0.02 SD lower for each 10 percentage point increase in the fraction of peers who are black (though they also find compositional peer effects to be highly nonlinear). Assuming the change in the composition of DC students' peers was on average similar to the overall change in the composition of students districtwide, this suggests that peer effects could only explain about 1 point of the increase in scores for DC students during this period. Combining the peer effects and direct effects of changing student compositions, we estimate that together they might explain about forty percent of the unadjusted impact estimates for grade 8 math shown in column 1 of Table V.2.

Table V.3. Estimated impacts on NAEP reading scores, by student race

	All Students		Subgroups		
	No controls	Demographic controls	Black	Hispanic	White
Grade 8 Reading					
Impact (NAEP scaled score)	3.2	1.0	-0.4	-7.1	n.a.
RMSPE	2.8	3.0	3.9	4.2	
Fit-adjusted *p*-value	0.19	0.68	0.96	0.05	
Effect size	0.09	0.03	-0.01	-0.20	
N (control states)	36	36	27	20	
Grade 4 Reading					
Impact (NAEP scaled score)	12.2*	8.8*	7.6*	8.1	0.1
RMSPE	3.1	3.0	5.0	5.1	2.9
Fit-adjusted *p*-value	0.03	0.03	0.03	0.09	0.76
Effect size	0.32	0.23	0.20	0.21	0.00
N (control states)	38	38	30	22	37

Source: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data, regression adjusted for student race and ethnicity in column 2.

Notes: RMSPE units are NAEP scaled score points. Subgroup reading scores for white grade 8 students in DC were not available via NCES because reporting standards were not met for this group in 1998, 2002, 2003, 2007, or 2009.

* p-value < 0.05.

n.a. = not available; NAEP = National Assessment of Educational Progress; RMSPE = root mean squared prediction error.

Estimated impacts are generally within one point of each other for black and Hispanic students with one exception—we estimate an impact of about -7 points on reading scores for Hispanic students in grade 8. Although this is a small subset of students in DC, the possibility that reforms may have had unintended negative consequences for this subgroup does suggest some cause for concern.

For grade 4, the estimated math impacts for white students are smaller than for other subgroups, at roughly 6 points for math, and essentially 0 points for reading. One possible explanation for the smaller impacts for white students is that a relatively higher proportion of them may have been attending higher quality schools in DC before the implementation of reforms. Separate estimates for white students in grade 8 are not reported because scores for this group were not available for DC for some years in the NCES data that report scores separately by state and student race and ethnicity. See notes in Table V.2 for details.

C. Estimates by years of exposure to reforms

The estimated impacts on grade 8 math achievement incorporate improvements in the education system from grades kindergarten through grade 8 and generally increase with the number of years a NAEP cohort was exposed to the PERAA reforms (Table V.4). The estimated impacts in reading also increase but remain much smaller than in math (Table V.5). The estimated impacts in math start at 4 points in 2009 for

the cohort that had only two years of exposure, and rise to at least 9 points by 2011 for the cohort with four years of exposure. The estimated impacts after 2011 vary between 9 and 10 points, depending on the model and year considered. Thus, it does appear that estimated impacts in 2009 are much smaller than in later years, a result that is consistent with much of the earlier literature on this topic. Similar results are seen in grade 4 math (Figure V.1). Estimated impacts on grade 8 reading are much less positive than on grade 8 math in all years, regardless of the method used.

As a result of these impacts over the past decade, DC has begun to catch up to the rest of the nation in terms of NAEP achievement levels. Indeed, as shown in Figure V.3, although DC has historically ranked last in state NAEP achievement, the impacts estimated here represent a closing of the gap between DC and other states. As of 2019, average NAEP performance in DC for grade 8 math was just narrowly above that of New Mexico and Alabama.

Table V.4. Grade 8 math impacts, by year

Grade 8 cohort (spring)	K–8 grades attended after PERAA	Years of PERAA exposure	Adjusting for student characteristics	
			No	Yes
2009	7–8	2	4.1	3.9
2011	5–8	4	10.2*	8.7*
2013	3–8	6	14.6*	12.4*
2015	1–8	8	13.6*	10.3*
2017	PK–8	10	16.6*	11.9*
Avg. impact			11.8*	9.4*
Avg. impact effect size			0.30*	0.24*
N (states)			**35**	**35**

Source: Authors' estimates using state-level NAEP data, regression adjusted for student demographics.

Notes: Table shows estimated average treatment effects for the treated. The average impact in the last column is the same as in Table V.1.

* *p*-value < 0.05.

Table V.5. Grade 8 reading impacts, by year

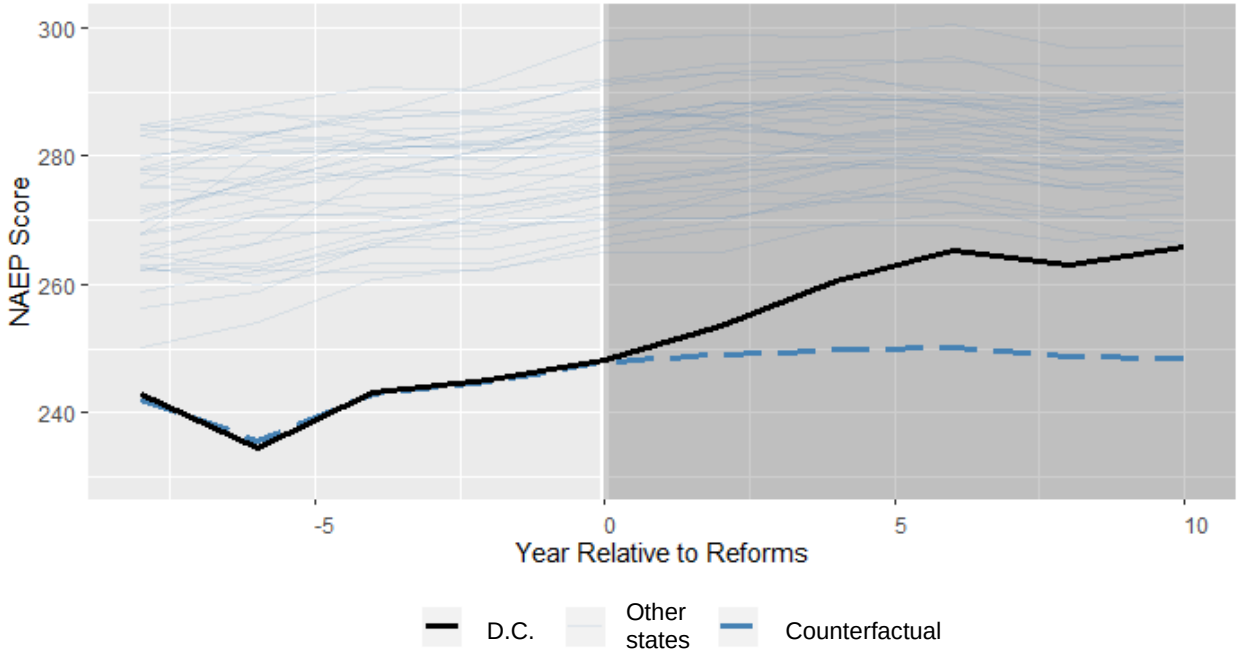
Grade 8 cohort (spring)	K–8 grades attended after PERAA	Years of PERAA exposure	Adjusting for student characteristics	
			No	Yes
2009	7–8	2	1.6	1.0
2011	5–8	4	0.1	-1.3
2013	3–8	6	3.8	1.6
2015	1–8	8	6.1*	3.2
2017	PK–8	10	4.3	0.5
Avg. impact			3.2	1.0
Avg. impact effect size			0.09	0.03
N (states)			**36**	**36**

Source: Authors' estimates using state-level NAEP data, regression adjusted for student demographics.

Notes: Table shows estimated average treatment effects for the treated. The average impact in the last column is the same as in Table V.1.

* *p*-value < 0.05.

Figure V.3. Grade 8 math scores, by year: DC, counterfactual, and other states



Source: Authors' estimates using state-level data on NAEP scores for grade 8 math.

VI. Conclusions

The reforms enacted in Washington, DC, beginning with PERAA in 2007 were an attempt to move beyond what was accomplished under NCLB by pushing for more accountability and more choices within public education. Not surprisingly, there has been a great deal of interest in understanding how successful the DC reform efforts were. Attempts to rigorously estimate impacts of these reforms have been hampered by the fact that DC is both a city and a state, rather than a city in a state containing other cities that used the same assessment. For these reasons, it is difficult to use standard evaluation methods to estimate impacts of city-level reforms there. We use state of the art methods to address these issues, focusing particularly on student achievement measured by the NAEP. We build on past efforts using an approach that could be easily adapted to study analogous effects in other cities. In particular, we use more years of data (covering roughly a quarter century) and recent advances in counterfactual estimation methods to more credibly account for changes over time in unobserved factors that influence achievement.

Our findings suggest the reforms DC implemented contributed to improvements for student achievement in math for both grade 4 students and grade 8 students. Indeed, relative to their counterparts in other areas, NAEP grade 4 math achievement among DC students appeared to improve by nearly one-third of a standard deviation compared to counterfactual outcomes constructed from similar geographic areas, after adjusting for both student characteristics and pre-reform trends in their outcomes. The results were also encouraging for grade 4 reading, where similar improvements were observed. The results in grade 8 reading were less strong, suggesting that the grade 4 improvements in that subject may have been covering skills that students would have picked up later in their education in the absence of the reforms. For math in grades 4 and 8, we observe larger estimated impacts for the later cohorts. This is consistent with the fact that the later cohorts were exposed to more years of the reforms that were implemented earlier and to additional reforms (e.g., IMPACT, added in 2009, and unified enrollment, added in 2014). The more positive impacts for math relative to reading are consistent with other evidence that schools and teachers may have more influence on math than reading achievement (Jacob 2005; Nye et al. 2004; Rivkin et al. 2005; Rockoff 2004).

At one-third of a standard deviation for math, the impacts we estimate for DC reforms are larger than those found for many education interventions such as class size reductions in Tennessee (0.19 SD) or the Success for All school reform program (0.11 SD) but also smaller than the impacts of some early childhood interventions like the Perry Preschool (0.49 SD) or the Abecedarian Project Preschool (0.65 SD). Thus, our estimates are within the range of what has been found elsewhere for math (Borman and Hewes 2002).

The impacts we find in DC are similar in magnitude to those observed by Harris and Larsen (2019) in New Orleans, where major school reforms were implemented starting in 2006–2007, immediately after hurricane Katrina. They report impacts between 0.28 and 0.40 SD on achievement in math, reading, science, and social studies. They also find that impacts increased over time during the years covered by their data (through 2014). This is consistent with our findings for DC because students in both cities had greater exposure to the reforms in the later years, and because both cities sequentially implemented more components of the reforms over time.

Our results in the early years of the DC reforms are also consistent with early year impacts found in Newark, where similar reforms were implemented. Those impacts were considerably smaller than our

overall estimated impacts in DC—around 0.08 SD in reading and no clear impacts in math (Chin et al. 2017). However, that study estimated impacts of the reforms in Newark over fewer years (2011–2012 through 2015–2016 school years). During that time, public charter school enrollment in Newark increased from 14 to 28 percent of the student population and other interventions related to teacher evaluations, curriculum reform, and turnaround schools were also implemented. Similarly, our estimated impacts for DC were much smaller in 2011 (four years after the interventions began) than they were in the later years of our data, which went through to the 2017–2018 school year.[19]

We also sought to estimate impacts on SAT scores, but were unable to produce credible impact estimates, due to changes in the percentage of students taking the SAT. For the interested reader, these analyses and results are discussed in Section III of Appendix A.

In comparison to other studies, our work brings in far more years of data, from both earlier and later years. This enables us to obtain a much richer picture of changes in achievement over time for DC students. We build on this strength by using methods designed to estimate impacts when there is a single treatment unit (DC in our case). Our results share similarities with other studies. Like NRC (2011) and Weiss and Long (2013) we find very small impacts immediately after the reforms were implemented. Like Özek (2014), Blagg and Chingos (2016), Carnoy et al. (2015), and Osborne and Langhorne (2018), we find impacts growing to be substantially larger in later years. In contrast to education interventions that show large initial impacts and fade out quickly (Kraft 2019), these results are quite encouraging in that the impacts in math both survive the introduction of new cohorts and compound over time. Although the methods we use require moderately strong assumptions regarding our measures of uncertainty, the results provide the most comprehensive evidence to date on a causal connection between the overall set of reforms enacted in DC and student achievement.

[19] Denver Public Schools also implemented major reforms in recent years and experienced large improvements in academic performance for their students (Baxter et al. 2019). However, we have not been able to find a careful study of results in Denver designed to help distinguish impacts of those reforms from other changes that might have been taking place during that time.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105(490): 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. Comparative politics and the synthetic control method. *American Journal of Political Science* 59(2): 495-510.

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff. 2017. Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis* 39(1): 54–76.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic difference in differences. No. w25532. National Bureau of Economic Research, 2019.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2018. Matrix completion methods for causal panel data models. NBER Working Paper No. 25132.

Bai, Jushan. 2009. Panel data models with interactive fixed effects. *Econometrica* 77(4): 1229–1279.

Blagg, Kristin, and Matthew Chingos. 2016. Does gentrification explain rising student scores in Washington, DC? Available https://www.urban.org/urban-wire/does-gentrification-explain-rising-student-scores-washington-dc. Accessed 21 December 2019.

Bohrnstedt, G., S, Kitmitto, B. Ogut, D. Sherman, and D. Chan. 2015. School Composition and the Black–White Achievement Gap (NCES 2015-018). U.S. Department of Education, Washington, DC: National Center for Education Statistics. Available at https://nces.ed.gov/pubsearch. Accessed March 3, 2020.

Borman, Geoffrey D., and Gina M. Hewes. 2002. The long-term effects and cost-effectiveness of Success for All. *Educational Evaluation and Policy Analysis* 24(4): 243–266. Available at https://journals.sagepub.com/doi/pdf/10.3102/01623737024004243. Accessed March 4, 2020.

Camilli, Gregory, Kentaro Yamamoto, and Ming-mei Wang. 1993. Scale shrinkage in vertical equating. *Applied Psychological Measurement* 17(4): 379–388.

Carnoy, Martin, Emma Garcia, and Tatiana Khavenson. 2015. Bringing It Back Home: Why State Comparisons Are More Useful Than International Comparisons for Improving U.S. Education Policy. EPI briefing paper #410. Washington, DC: Economic Policy Institute.

Chingos, Matthew, Kristin Blagg, and Grace Luetmer. 2019. *America's Gradebook: How does your state stack up? Appendix.* Washington, DC: The Urban Institute. Available http://apps.urban.org/features/naep/naep-technical-appendix.pdf. Accessed 21 December 2019.

Christie, Kathy, Michael Colasanti, and Dinah Frey. 2010. *State teacher tenure/continuing contract laws.* Denver, CO: Education Commission of the States. Available https://www.ecs.org/clearinghouse/88/28/8828.pdf. Accessed 21 December 2019.

College Board. 2007. *2007 College-Bound Seniors, State profile Report Dist of Columbia*. The College Board. New York, NY. Available at https://secure-media.collegeboard.org/digitalServices/pdf/research/cb-seniors-2007-DC.pdf. Accessed 1/3/2020.

District of Columbia (DC). 2006. Out of boundaries transfers rule. 53 DCR 9195. Available https://www.dcregs.dc.gov/Co mmon/DCMR/RuleDetail.aspx?RuleId=R0024934. Accessed 21 December 2019.

Doudchenko, Nikolay, and Guido W. Imbens. 2016. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. NBER Working Paper No. 22791.

Education Consortium for Research and Evaluation. 2014. *A closer look at student achievement trends in the District of Columbia: School years 2006–2007 and 2012–2013*. DC PERAA Report No. 4. Washington, DC: Office of the District of Columbia Auditor.

Gitomer, D. H., K. Crouse, and Jeanette Joyce. 2014. *A review of the DC IMPACT teacher evaluation system.* New Brunswick, NJ: Graduate School of Education, Rutgers University.

Gobillon, Laurent, and Thierry Magnac. 2016. Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics* 98(3): 535–551.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 2009. "New evidence about Brown v. Board of Education: The complex effects of school racial composition on chievement." *Journal of Labor Economics* 27(3): 349–383.

Harris, Douglas N., and Matthew F. Larsen. 2019. The effects of the New Orleans post-Katrina market-based school reforms on student achievement, high school graduation, and college outcomes. New Orleans, LA: Education Research Alliance for New Orleans.

Hastie, Trevor, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. 2015. Matrix completion and low-rank SVD via fast alternating least squares. *The Journal of Machine Learning Research* 16 (1): 3367–3402.

Hesla, Kevin. 2018. *Unified enrollment lessons learned from across the country*. Washington, DC: National Alliance for Public Charter Schools. Available at https://www.publiccharters.org/sites/default/files/documents/2018-09/rd3_unified_enrollment_web.pdf. Accessed 21 December 2019.

Hewins, Jessie, Randy Rosso, and Alison Maurice. 2017. *Community Eligibility Continues to Grow in the 2016–2017 School Year*. Washington, DC: Food Research and Action Center. Available at https://www.frac.org/wp-content/uploads/CEP-Report_Final_Links_032317.pdf. Accessed on 12/31/2019.

Hurst, David, Alexandra Tan, Anne Meek, and Jason Sellers. 2008. *Overview and inventory of state education reforms: 1990 to 2000.* NCES Report No. 2003–020. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Jacob, B. A. 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5-6), 761-796.

Jacobs, Leslie. 2015. New Orleans by the numbers: Public school enrollment. Available https://educatenow.net/2015/01/28/new-orleans-by-the-numbers-public-school-enrollment/. Accessed 21 December 2019.

National Alliance for Public Charter Schools. 2014. *A growing movement: America's largest charter school communities* (9th ed.). Washington, DC: National Alliance for Public Charter Schools. Available http://www.publiccharters.org/sites/default/files/migrated/wp-content/uploads/2014/12/2014_Enrollment_Share_FINAL.pdf. Accessed 21 December 2019.

National Center for Education Statistics. 2010. *The Nation's Report Card: Grade 12 Reading and Mathematics 2009 National and Pilot State Results*. NCES 2011–455. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Center for Education Statistics. 2017. NAEP Assessment Sample Design. Available at https://nces.ed.gov/nationsreportcard/tdw/sample_design/. Accessed 27 December 2019.

National Center for Education Statistics. 2018a. Table V.221.75. Average National Assessment of Educational Progress (NAEP) reading scale score and standard deviation, by selected student characteristics, percentile, and grade: Selected years, 1992 through 2017. *Digest of Education Statistics*. Available https://nces.ed.gov/programs/digest/d18/tables/dt18_221.75.asp. Accessed 27 December 2019.

National Center for Education Statistics. 2018b. Trial Urban District Assessment (TUDA). Available https://nces.ed.gov/nationsreportcard/tuda/. Accessed 21 December 2019.

National Center for Education Statistics. 2019. Fast facts. Available https://nces.ed.gov/fastfacts/display.asp?id=514. Accessed 27 December 2019.

National Research Council, Committee on the Independent Evaluation of DC Public Schools, Division of Behavioral and Social Sciences and Education (NRC). 2011. *A plan for evaluating the District of Columbia's Public Schools: From impressions to evidence.* Washington, DC: The National Academies Press.

National Research Council (NRC). 2015. *An Evaluation of the Public Schools of the District of Columbia: Reform in a Changing Landscape. Committee for the Five-Year (2009-2013) Summative Evaluation of the District of Columbia's Public Schools. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education.* Washington, DC: The National Academies Press. Available at http://zd4l62ki6k620lqb52h9ldm1.wpengine.netdna-cdn.com/wp-content/uploads/2018/12/An-Evaluation-of-the-Public-Schools-of-the-District-of-Columbia-Reform-in-a-Changing-Landscape.pdf.

Nye, Barbara, Spyros Konstantopoulos and Larry V. Hedges. 2004. "How Large Are Teacher Effects?" *Educational Evaluation and Policy Analysis,* 2004, 26 (3): 237-257. Available at https://www.sesp.northwestern.edu/docs/publications/169468047044fcbd1360b55.pdf.

Osborne, David and Emily Langhorne. 2018. Analysis: NAEP scores show D.C. is a leader in educational improvement—with powerful lessons for other cities [Blog post]. Available https://www.the74million.org/article/analysis-naep-scores-show-d-c-is-a-leader-in-educational-improvement-with-powerful-lessons-for-other-cities/. Accessed 21 December 2019.

Özek, Umut. 2014. A closer look at the student achievement trends in the District of Columbia between 2006–07 and 2012–13. Working Paper No. 119. Washington, DC: American Institutes for Research. Available https://eric.ed.gov/?id=ED553414. Accessed on 21 December 2019.

Reardon, Sean F., Joseph Cimpian, and Ericka S. Weathers. 2014. Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In *Handbook of Research in Education Finance and Policy, Second Edition*, pp. 491–509. Taylor and Francis.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. 2005. Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417-458.

Rockoff, J. E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2), 247-252.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley.

Walsh, Elias, and Dallas Dotter. 2014 *The impact of replacing principals on student achievement in DC public schools.* Princeton, NJ: Mathematica Policy Research.

Weiss, E., and D. Long. 2013. Market-oriented education reforms' rhetoric trumps reality: The impacts of test-based teacher evaluations, school closures, and increased charter school access on student outcomes in Chicago, New York City, and Washington, DC. Washington, DC: Broader, Bolder Approach to Education. Available http://www.epi.org/files/2013/bba-rhetoric-trumps-reality.pdf. Accessed 12 October 2017.

Xu, Yiqing. 2017. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* 25(1): 57–76.

Appendix A: Supplemental Analyses

A. Sensitivity analyses

1. Choice of geographic units

Our estimates using counties as the geographic unit of analysis are similar to those that use states, with slightly larger estimates using counties (Table A.1). In the main body of the report we focus on the state level results because they are more precise based on the root mean squared prediction error (RMSPE). The reported p-values are derived from the empirical distribution of results across all units. Consequently, similar p-values between estimates does not equate to similar precision of estimates or relative size of standard errors. When looking at estimation error, the estimates using county-level scores are subject to larger RMSPEs than those using state-level scores (roughly 1.7 times larger for math and 2 times larger for reading). This is, in part, due to far fewer schools contributing to county averages than state averages and the fact that different schools are sampled each year of NAEP. Hence within-county average NAEP scores exhibit more year-to-year fluctuations due to sampling error than within-state averages. The lower RMSPE for state-level placebo tests suggests our methods can construct more accurate counterfactuals using state-level data than using county-level data.

As with the state-level analysis, we do not find strong evidence of a positive impact on grade 8 reading. For the other grade and subject combinations, the county estimates are roughly 1 to 4 NAEP scaled score points higher than the corresponding state estimates.

Table A.1. NAEP impact estimate comparisons: states versus counties

	Grade 8 Math	Grade 4 Math	Grade 8 Reading	Grade 4 Reading
States				
Impact (NAEP scaled score)	11.7*	11.1*	3.4	9.8*
RMSPE	3.0	2.9	2.6	2.8
Fit-adjusted *p*-value	0.03	0.03	0.11	0.03
N (states)	35	35	36	38
Effect size (SD units)	**0.30**	**0.36**	**0.09**	**0.26**
Counties				
Impact (NAEP scaled score)	12.6*	13.7*	2.8	13.4*
RMSPE	5.2	4.7	5.3	5.6
Fit-adjusted *p*-value	0.03	0.02	0.45	0.03
N (counties)	40	60	50	60
Effect size (SD units)	0.32	0.44	0.08	0.35

Source: State estimates: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data. County estimates: Authors' estimates using county-level NAEP data, 1996–2017 Mathematics and 1998–2017 Reading Assessments.

Notes: Impacts and RMSPE units are NAEP scaled score points. County sample sizes rounded to the nearest 10 to protect confidentiality.

* *p*-value < 0.05.

RMSPE = root mean squared prediction error; SD = standard deviation.

2. Capturing grade 4 to 8 growth

In the main body of the report we focus on estimating impacts on grades 4 and 8 performance separately. In this section we estimate impacts on within-cohort gains by looking at performance in grade 8 minus grade 4 from the same cohort, four years earlier. Estimated impacts on within-cohort gains in NAEP achievement are small in math and negative (though not statistically significant at the 0.05 level) in reading (Table A.2). They are similar regardless of whether we use the state or county data. One interpretation of these results is that our grade 4 and 8 estimates are biased because of unobserved factors. Another interpretation, however, is that most of the growth in math skills occurred by grade 4 and that there was relatively little improvement between grades 4 and 8 in math. For reading, it appears that any benefits of the reform achieved by grade 4 disappeared by grade 8. The latter is consistent with the lower estimated impacts on growth for students in later cohorts, which had the most exposure to the reforms and saw the greatest impacts on NAEP achievement in grade 4.

Table A.2. NAEP gains impact estimate comparisons: states versus counties

	Grade 4-8 Math Gains	Grade 4-8 Reading Gains
States		
Impact (NAEP scaled score)	3.6	-7.6
RMSPE	3.0	6.7
Fit-adjusted *p*-value	0.06	0.06
N (states)	30	28
Effect size (SD units)	**0.09**	0.20
Counties		
Impact (NAEP scaled score)	3.7	-7.7
RMSPE	5.9	7.5
Fit-adjusted *p*-value	0.36	0.06
N (counties)	30	50
Effect size (SD units)	0.09	0.22

Source: State estimates: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data. County estimates: Authors' estimates using county-level NAEP data, 1996–2017 Mathematics and 1998–2017 Reading Assessments.

Notes: Impacts and RMSPE units are NAEP scaled score points. County sample sizes rounded to the nearest 10 to protect confidentiality.

* *p*-value < 0.05.

RMSPE = root mean squared prediction error; SD = standard deviation.

3. Limiting the comparison units to exclude those using similar reforms

The overall pattern of the main results in Table V.1 also holds regardless of whether all units in the sample are available to contribute to counterfactual estimates, or whether the analytic sample is restricted to only those states or counties that had not implemented reforms similar to those in DC during the periods of interest. Estimates using state-level data are nearly identical whether using the full set of states with balanced panel data or excluding the four states among that group that had implemented policies of no teacher tenure at some point between 2007 and 2014 (Table A.3). Similarly, estimates using all counties in our sample are nearly identical to estimates using only the counties that had not implemented similar reforms (Table A.4). That estimates are quite similar regardless of whether states, counties, or restricted samples of either are used is perhaps not surprising for two reasons. First, few states and counties are removed when we exclude those that we determined had implemented reforms similar to those in DC. Second, the methods we use to estimate counterfactual outcomes are flexible in how they use variation between units and across years to best approximate the treated unit. It is not difficult for these methods to find weighted combinations across those two dimensions that produce similar results after minor changes to the units included in the data.

We also present our impacts as effect sizes in Table A.1, using standard deviations of the outcomes as units. Impacts are often presented in effect sizes given uncertainty about how to interpret scale scores. However, we present our results using scale scores in the main body of the report and in all subsequent tables and figures in our appendices because the NAEP scores were designed originally to be compared

across grade levels (Camilli et al. 1993). While it is difficult to design tests for this purpose, we believe that it is still preferable to use the scale scores when doing cross-grade comparisons rather than the effect size units which are clearly not designed for this purpose since the standard deviations of outcomes can change across grades.

Table A.3. NAEP average treatment comparisons: all states versus restricted sample

	Grade 8 Math	Grade 4 Math	Grade 8 Reading	Grade 4 Reading
All States				
Impact (NAEP scaled score)	11.7*	11.1	3.4	9.8
RMSPE	3.0	2.9	2.6	2.8
Fit-adjusted *p*-value	0.03	0.03	0.11	0.03
N (states)	35	35	36	38
States Without Reforms				
Impact (NAEP scaled score)	11.5	10.9	3.6	10.1
RMSPE	3.1	2.9	2.6	2.8
Fit-adjusted *p*-value	0.03	0.03	0.09	0.03
N (states)	33	33	34	35

Source: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data, regression adjusted for student demographics.

Notes: Impacts and RMSPE units are NAEP scaled score points.

* *p*-value < 0.05.

Table A.4. NAEP average treatment comparisons: all counties versus restricted sample

	Grade 8 Math	Grade 4 Math	Grade 8 Reading	Grade 4 Reading
All Counties				
Impact (NAEP scaled score)	12.6*	13.7*	2.8	13.4*
RMSPE	5.2	4.7	5.3	5.6
Fit-adjusted *p*-value	0.03	0.02	0.45	0.03
N (counties)	40	60	50	60
Counties Without Reforms				
Impact (NAEP scaled score)	12.6*	13.6*	2.9	13.9*
RMSPE	5.2	4.9	5.1	5.5
Fit-adjusted *p*-value	0.03	0.02	0.42	0.03
N (counties)	40	50	40	60

Source: Authors' estimates using county-level National Assessment of Educational Progress (NAEP) data, 1996–2017 Mathematics and 1998–2017 Reading Assessments.

Notes: Impacts and RMSPE units are NAEP scaled score points. Sample sizes rounded to the nearest 10 to protect confidentiality.

* *p*-value < 0.05.

RMSPE = root mean squared prediction error.

B. DC student compositions: 1990–2017

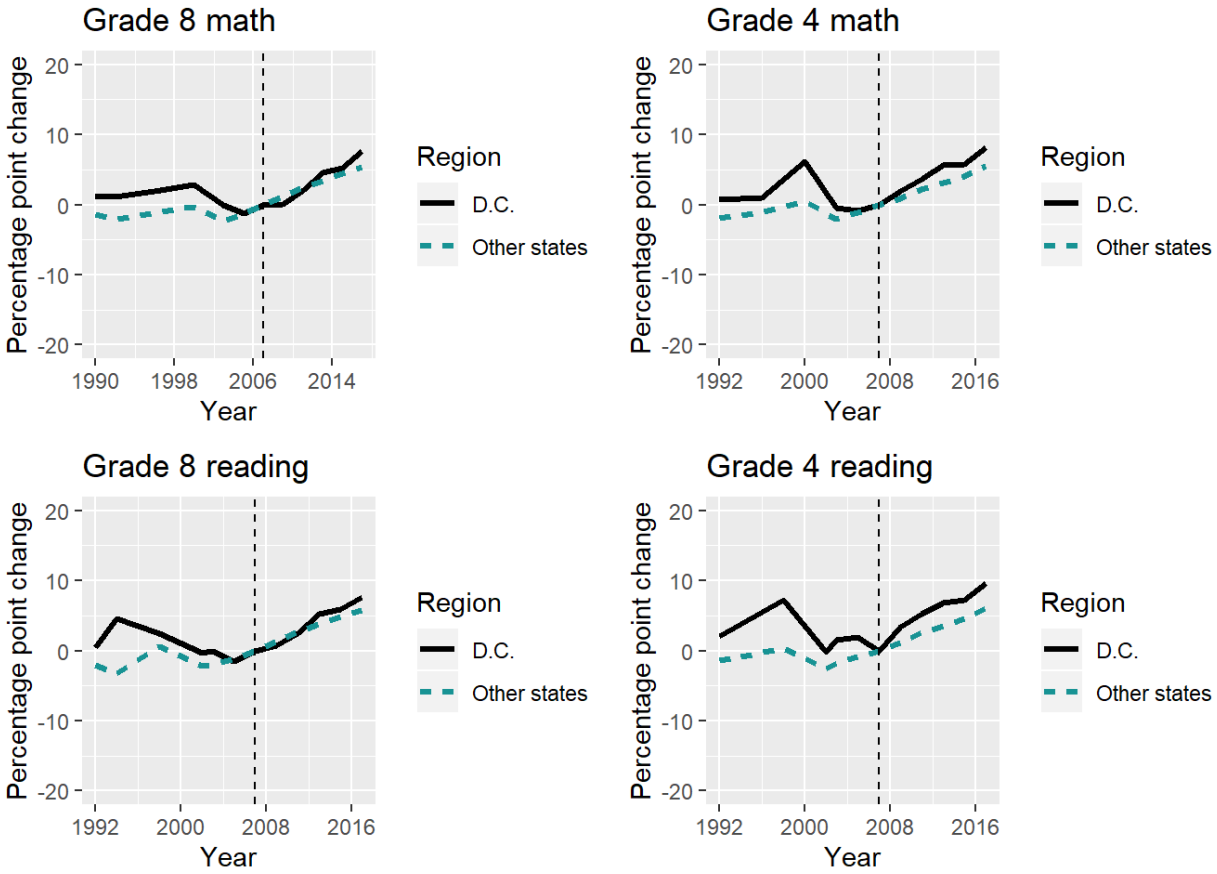
The percentages of students in DC who are black and white have changed substantially over time and are reflected in the composition of students sampled by NAEP each year. In particular, the percentage of students who are white has risen, whereas the percentage who are black has fallen, both by about 15 percentage points since 2007, relative to their percentages nationally. In contrast, there have been no clear changes, relative to national averages, in the percentages of DC students who are Hispanic or of other races and ethnicities (Figures A.1 and A.2). As might be expected, the results are quite similar among grade 4 and grade 8 students sampled by NAEP, for either reading and math. The trends before 2007 are more complex, with the percentages bouncing up and down between 1990 and 2007, involving transitory dips and spikes for Black and White students in DC, respectively. The pre-2007 changes may reflect changes in the definitions of racial and ethnic groups, given that the changes are seen in the same years in both grades 8 and 4.[20] However, the trends present between 2002 and 2017 indicate a steadily declining proportion of DC students who are black. For this reason, Tables V.2 and V.3 explore the extent to which changing student compositions might explain the effects reported in Table V.1 and Figure V.1.

As shown in Figure A.3, there is no clear difference between DC and the rest of the nation in the trends over time for the percentage of students eligible for free or reduced-price school meals, but the upward trend likely reflects that ways of coding that variable have changed over time and a growing number of schools treat all students as eligible, rather than a steady decrease in average household income levels among students (Hewins et al. 2017). Similarly, Figure A.4 shows no clear differences in the trend over time for the percentage of students identified as learning English as a second language (ESL). Figure A.5 shows that the percentage of students identified as enrolled in special education has risen by perhaps five to eight percentage points in DC relative to students nationwide. Thus, controlling for the percentage of students in special education might increase the estimated impacts over time.

It is not clear whether it is more appropriate to control for special education status. It is quite possible that, all else equal, black students receive fewer special education services than they should (Gordon 2017). If this is true, then efforts to increase the use of special education in DC might be part of a successful reform effort and controlling for special education status would underestimate impacts of those reforms.

[20] A similar pattern is found in the Common Core of Data (CCD), albeit in different years. If the changes were real, we would expect that the grade 8 changes would occur approximately four years after the grade 4 changes and that similar changes would be found in NAEP and CCD. Instead we see that the grade 4 and 8 changes occur in the same years in each dataset but in different years across datasets, suggesting that they reflect definitional changes and not true changes in the population. Percentage point shifts were larger in DC than elsewhere because DC has a much higher percentage of black students than the rest of the nation. These definitional changes should not bias our estimates given that in each year the same definitions are used in DC and in the comparison units.

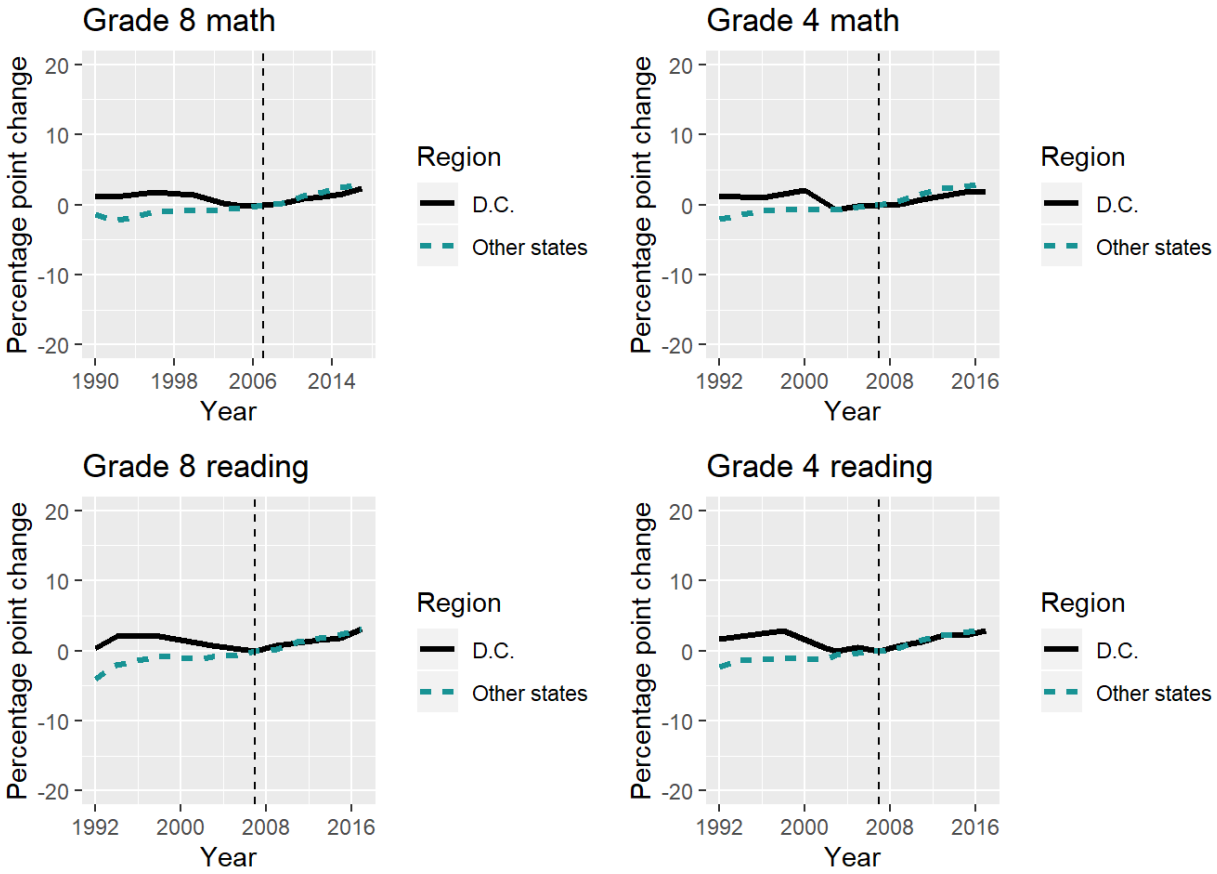
Figure A.1. Percentage of Hispanic students in DC and other states, relative to 2007



Source: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data.

Notes: These data cover the entire United States and not just the samples of counties and states used in our analyses. All percentages are set to the original values minus the value in 2007.

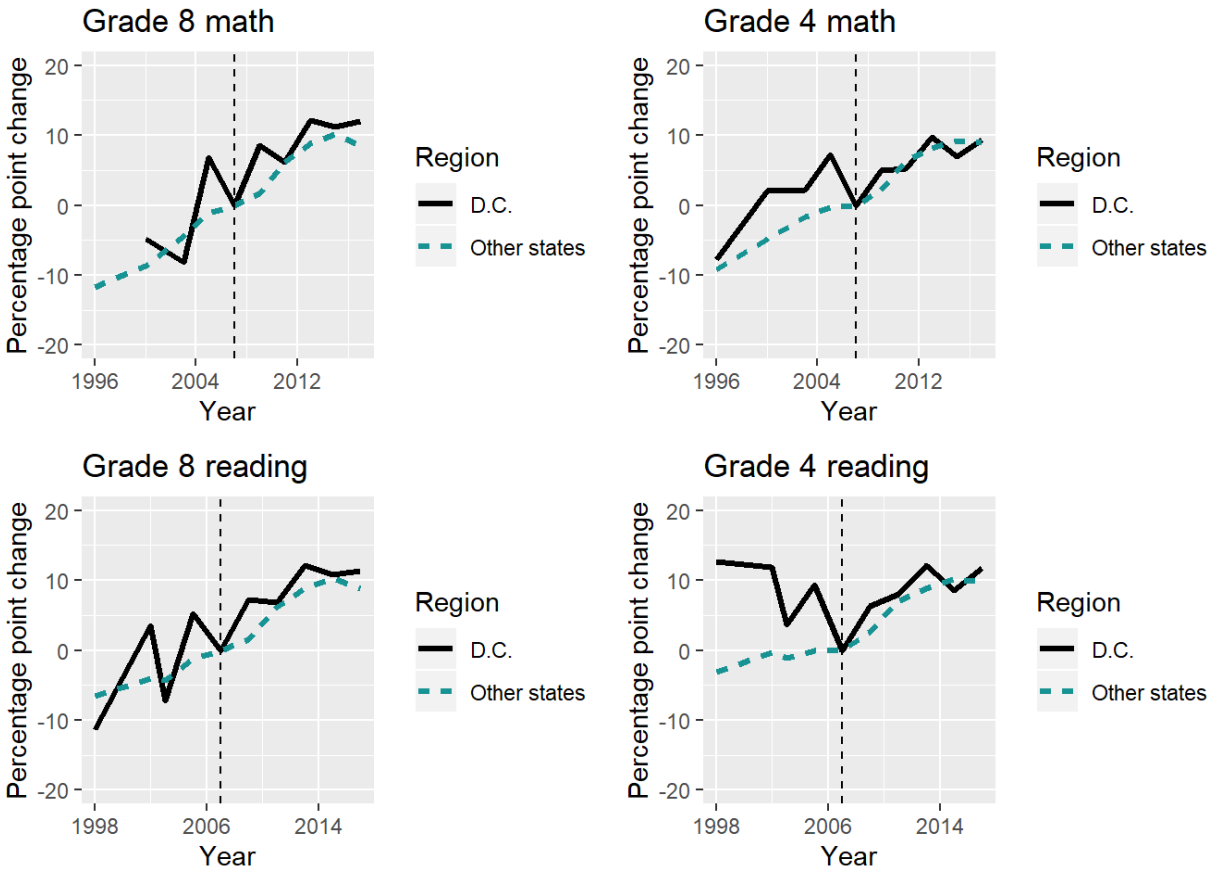
Figure A.2. Percentage of "other" race students in DC and in other states, relative to 2007



Source: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data.

Notes: These data cover the entire United States and not just the samples of counties and states used in our analyses. All percentages are set to the original values minus the value in 2007.

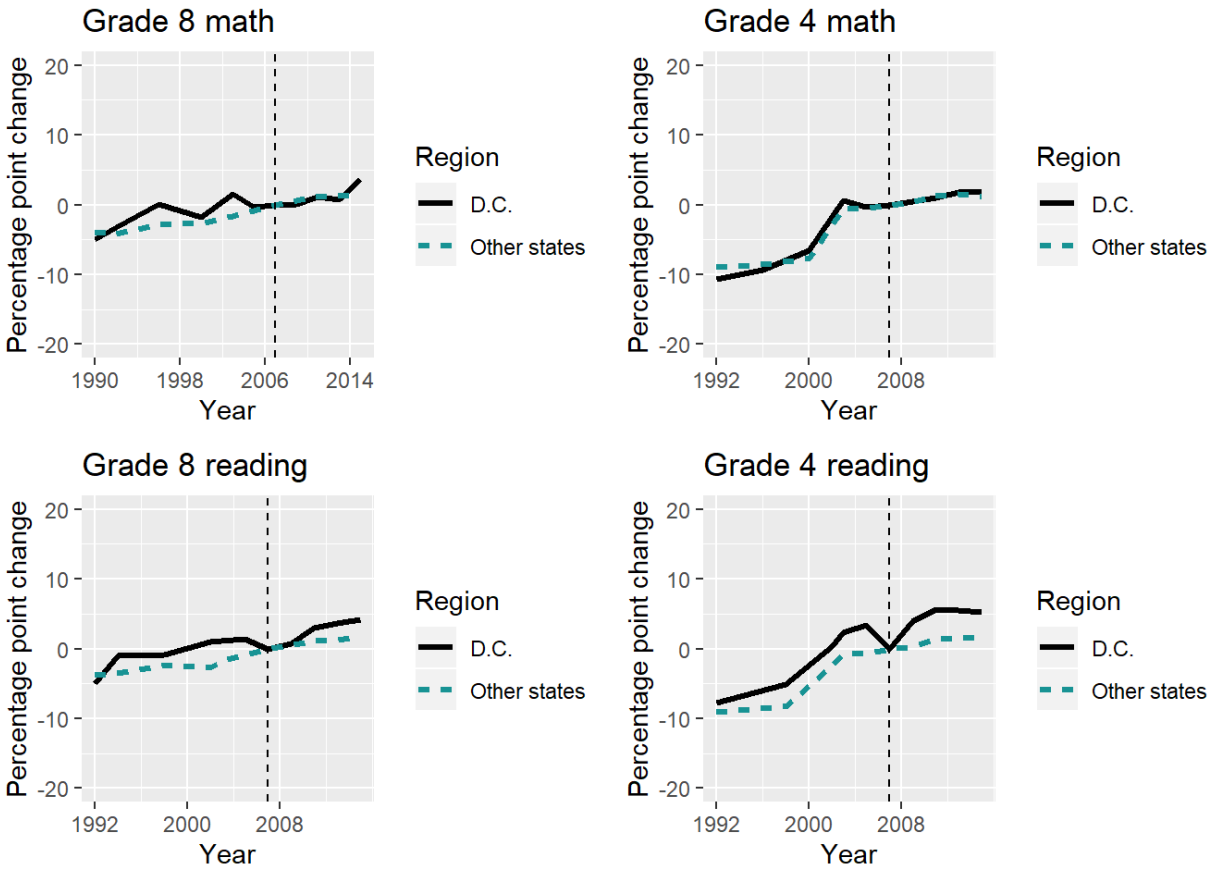
Figure A.3. Percentage of students eligible for free or reduced-price school meals in DC and other states, relative to 2007



Source: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data.

Notes: These data cover the entire United States and not just the samples of counties and states used in our analyses. All percentages are set to the original values minus the value in 2007.

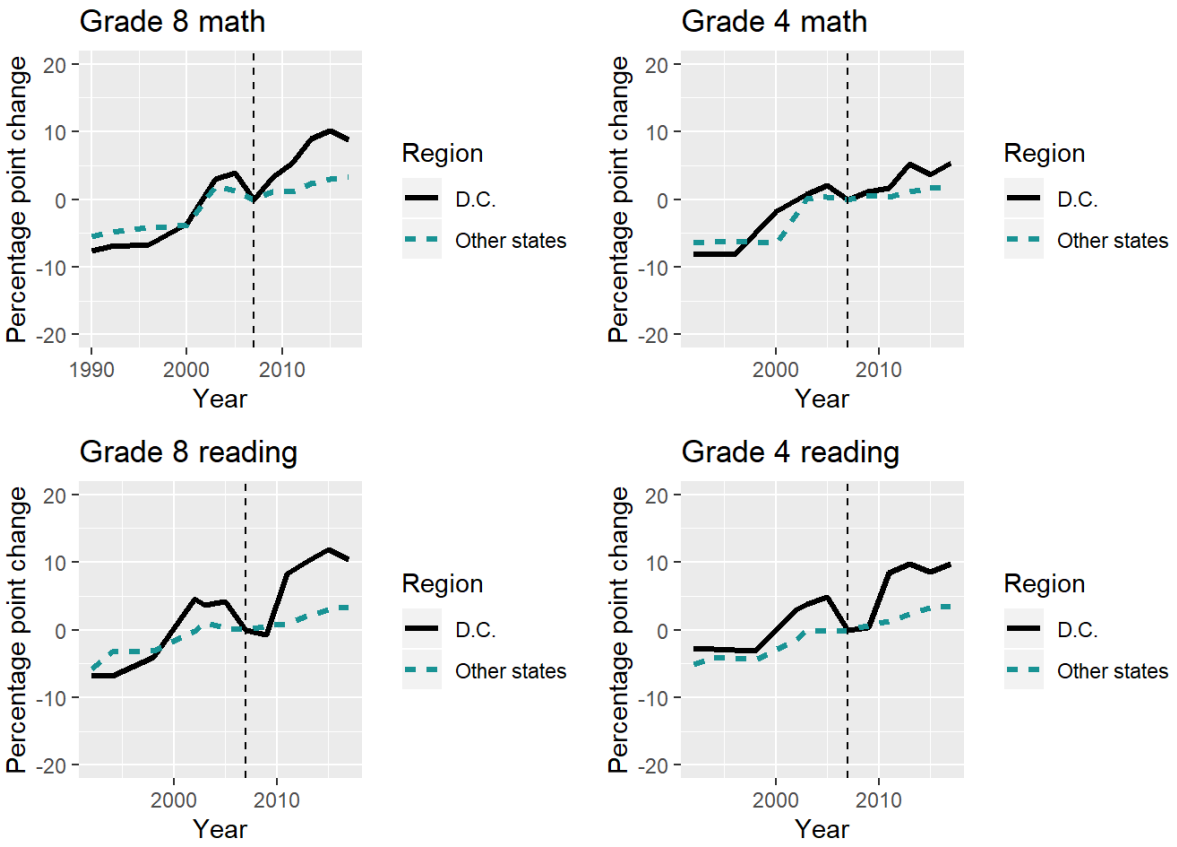
Figure A.4. Percentage of students with English as a second language in DC and other states, relative to 2007



Source: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data.

Notes: These data cover the entire United States and not just the samples of counties and states used in our analyses. All percentages are set to the original values minus the value in 2007.

Figure A.5. Percentage of students in special education in DC and other states, relative to 2007



Source: Authors' estimates using state-level National Assessment of Educational Progress (NAEP) data.

Notes: These data cover the entire United States and not just the samples of counties and states used in our analyses. All percentages are set to the original values minus the value in 2007.

C. Impacts on SAT participation and achievement

We use data from the College Board's "College-Bound Seniors" reports for SAT outcomes. These reports are produced annually for each state, including the District of Columbia, and present data for students who graduated from high school in the report year and had taken the SAT. According to the College Board, students who took the SAT more than once while in high school are counted only once, and only their latest scores are used for reporting.[21]

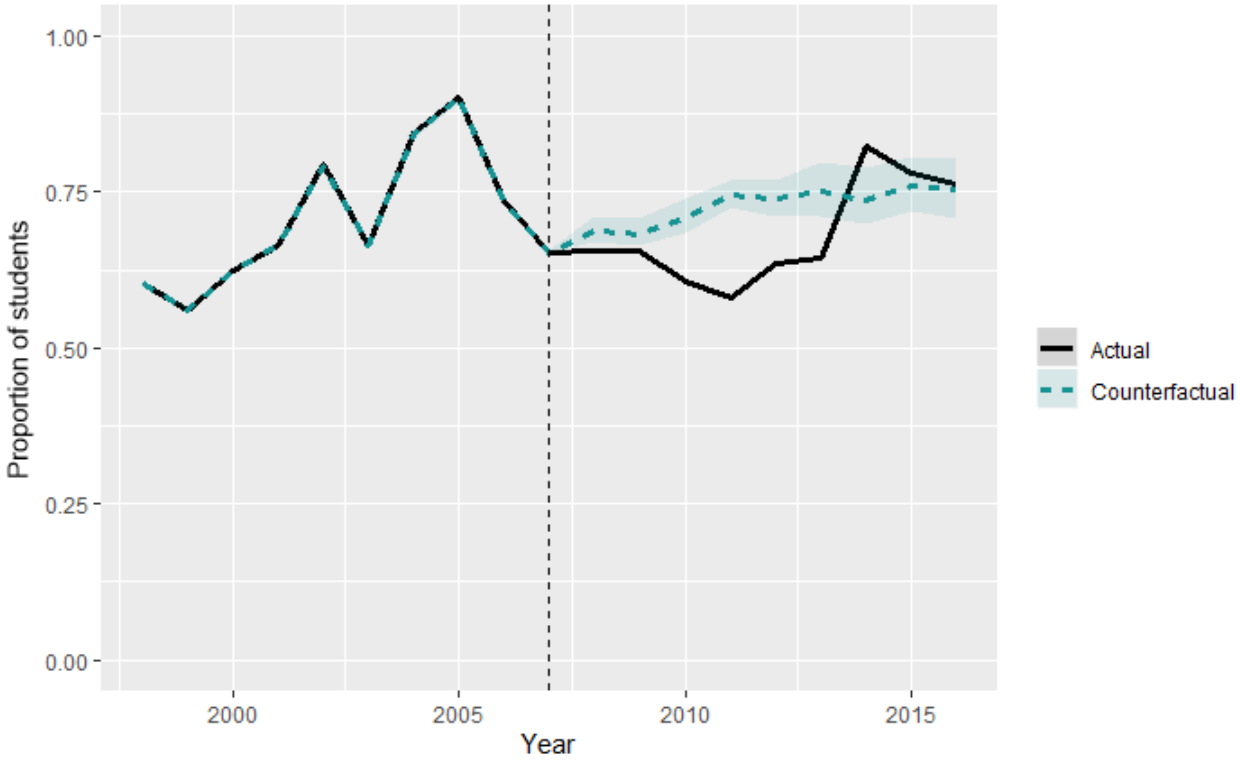
Using data elements from these reports, we construct panel data on several SAT measures for SAT takers in 51 states over 19 consecutive years (corresponding to high school graduates from the classes of 1998 through 2016). These include SAT score means and the number of test takers for each of the verbal, math, and writing portions of the test. We use the number of graduating seniors who took the SAT relative to the number of grade 9 students in the same cohort four years earlier to estimate the impacts for SAT participation, noting that this captures a combined outcome of SAT participation and high school graduation. We also analyze these measures separately for three sets of subgroups: student race and gender, household income level, and the highest education attainment level among the student's parents.[22]

Overall, we find no clear evidence that the reforms implemented in DC had impacts on SAT participation or achievement. We think the achievement results may be biased due to the relatively low and varying participation rates. Also we think it is quite plausible that there were minimal effects by grade 12 since it appears that most of the benefits of the reforms were in the early grades and the students who reached grade 12 in our sample had not had much exposure to these reforms before grade 4. We find initial negative impacts on participation after 2007, followed by a dramatic shift to a positive impact in 2014 (Figure A.6). This corresponds to a deliberate policy among DCPS to have all high school juniors take the SAT for free, in class during a school day, beginning with the class of 2014.

[21] See, for example, the 2007 College-Bound Seniors state profile report for the District of Columbia, available at https://secure-media.collegeboard.org/digitalServices/pdf/research/cb-seniors-2007-DC.pdf.

[22] The reporting of scores by annual household income is broken out by income ranges (for example, $10,000 ranges top-coded at $100,000 in the 2007 report), which are not adjusted for inflation. This prohibits a direct adjustment for inflation over time. Hence, we calculate nominal household income quantiles each year and choose groupings of the discrete income ranges that yield the most stable subgroup definition in terms of household income percentile.

Figure A.6. SAT participation in DC and estimated counterfactual



Source: Authors' estimates using state-level SAT and CCD data.

Notes: Actual = observed SAT math scores for DC (solid line). Counterfactual = matrix completion estimates of counterfactual outcomes in the absence of reforms (dashed line). Vertical dotted line demarcates the implementation of PERAA in 2007. Shaded region is the 95 percent confidence interval based on the post-2007 RMSPE for observed counterfactual SAT scores across all other states. The overall impact across years is shown in Table A.5.

RMSPE = root mean squared prediction error.

Table A.5. Estimated impact on SAT participation

	Impact (proportion of students)	RMSPE	Fit-adjusted *p*-value
SAT Participation	-0.05	0.09	0.14
N (states) = 50			

Source: Author's estimates using state-level SAT and CCD data.

RMSPE = root mean squared prediction error.

Most point estimates for impacts on SAT scores are negative but are not precisely estimated (Table A.6, Figures A.7 and A.8). The impacts are also smaller in magnitude compared to those estimated for NAEP scores. For example, although we estimate a positive effect of the reforms on grade 8 NAEP math achievement of about 0.3 SD (as discussed above), we estimate negative impacts on SAT math achievement by race and gender of about minus 0.1 SD or less.[23] While we focus on scale scores when describing NAEP results, we also report effect size units for the SAT and NAEP results because the SAT and NAEP scale scores were not designed to be compared with each other.

Table A.6. Estimated impacts on SAT scores, overall and by race and gender

		Female		Male	
	All	Black	White	Black	White
SAT Math Scores					
Impact (SAT score)	-18.5	-12.5	-7.2	-16.3	-0.3
RMSPE	14.7	18.6	11.4	20.8	14.4
Fit-adjusted *p*-value	0.14	0.59	0.25	0.30	0.92
N (states)	50	45	50	45	50
Effect size (SD units)	-0.12	-0.08	-0.05	-0.11	0.00
SAT Reading Scores					
Impact (SAT score)	-23.9	-20.4	-5.2	-23.8	-2.5
RMSPE	14.1	17.4	12.2	20.2	14.6
Fit-adjusted *p*-value	0.08	0.17	0.29	0.15	0.73
N (states)	50	45	50	45	50
Effect size (SD units)	-0.16	-0.14	-0.03	-0.16	-0.02

Source: Author's estimates using state-level SAT data.

Notes: RMSPE is calculated using prediction errors for post-2007 outcomes under individual placebo tests across states in the analytic sample. Fit-adjusted p-values are based on the samplewide distribution of the statistic described in our empirical strategy section, which is closely related to RMSPE and also model fit in the pre-treatment years.

RMSPE = root mean squared prediction error; SD = standard deviation.

[23] The standard deviations for the DC SAT scores in math and reading for 2007 were 150 and 149 points, respectively (College Board 2007).

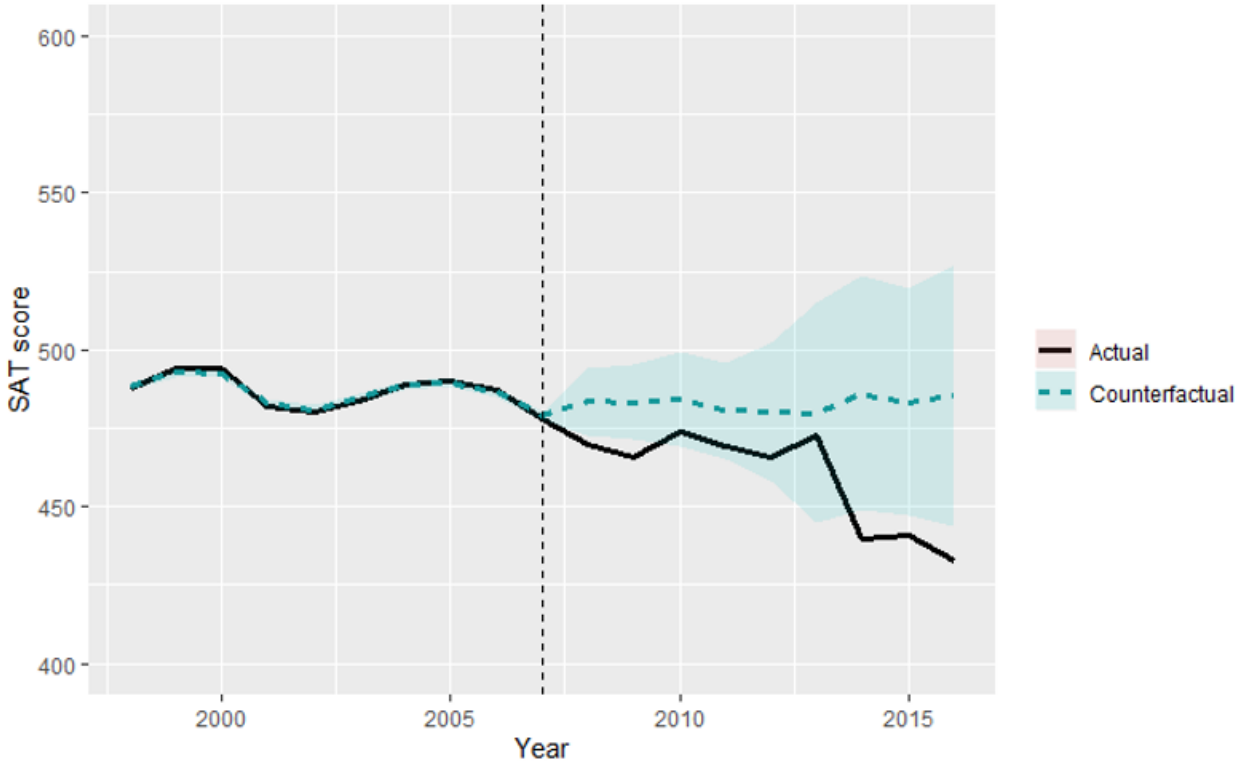
Figure A.7. SAT math scores for DC and estimated counterfactual

Source: Authors' estimates using state-level SAT data.

Notes: Actual = observed SAT math scores for DC (solid line). Counterfactual = matrix completion estimates of counterfactual outcomes in the absence of reforms (dashed horizontal line). Vertical dotted line demarcates the implementation of PERAA in 2007. Shaded region is the 95 percent confidence interval based on the post-2007 RMSPE for observed counterfactual SAT scores across all other states.

RMSPE = root mean squared prediction error.

Figure A.8. SAT reading scores for DC and estimated counterfactual



Source: Authors' estimates using state-level SAT data.

Notes: Actual = Observed SAT reading scores for DC (black line). Counterfactual = matrix completion estimates of counterfactual outcomes in the absence of reforms (dashed horizontal line). Vertical dotted line demarcates the implementation of PERRA in 2007. Shaded region is the 95 percent confidence interval based on the post-2007 RMSPE for observed counterfactual SAT scores across all other states.

Separating impacts on SAT scores by household income or parents' educational attainment reveals estimates that are smaller in magnitude for each subgroup, compared to overall impacts (Tables A.7 and A.8). This pattern of results appears to be explained by the changing composition of SAT test takers in DC over the post-treatment period. More precisely, this pattern suggests that the overall impacts on SAT scores may be biased downwards because the population of test takers changed to include more low-performing students. For example, the percentage of SAT test takers in DC among students from households with incomes below the federal poverty line rose from about 30 percent in 2007 to 47 percent in 2016. Consequently, although neither group experienced a meaningful impact, average SAT scores decreased over this time period as more students from low-income households participated in the SAT. A similar story may hold for parent education if students with lower levels of parent education increased their rates of taking the SAT by far more than students whose parents had higher levels of education—in part because those students may have already been taking the SAT at high rates.

In sum, we do not believe the data allow us to produce credibly unbiased estimates of the impacts of the reforms in DC on SAT participation and scores.

Table A.7. Estimated impacts on SAT scores, by parent educational attainment

	All	No HS	HS	AA	BA	Grad
SAT math scores						
Impact (SAT score)	-18.5	-10.9	0.2	4.0	0.8	-8.2
RMSPE	14.7	30.4	15.2	14.9	10.5	11.2
Fit-adjusted *p*-value	0.14	0.51	0.76	0.46	0.45	0.43
N (states)	50	46	50	49	50	50
Effect size (SD units)	-0.12	-0.07	0.00	0.03	0.01	-0.05
SAT verbal scores						
Impact (SAT score)	-23.9	-6.0	-4.9	-3.2	-1.6	-10.2
RMSPE	14.1	23.2	14.5	13.8	10.1	10.7
Fit-adjusted *p*-value	0.08	0.49	0.31	0.36	0.35	0.29
N (states)	50	46	50	49	50	50
Effect size (SD units)	-0.16	-0.04	-0.03	-0.02	-0.01	-0.07

Source: Authors' estimates using state-level SAT data.

RMSPE = root mean squared prediction error; SD = standard deviation; No HS = no high school degree; HS = high school degree but no post-secondary degree; AA = two-year degree but no BA; BA = four-year degree but no post-graduate degree; Grad = graduate degree.

Table A.8. Estimated impacts on SAT scores, by household income level

	All	Below FPL	Above FPL
SAT math scores			
Impact (SAT score)	-18.5	-3.2	3.0
RMSPE	14.7	10.8	14.5
Fit-adjusted *p*-value	0.14	0.31	0.68
N (states)	50	47	47
Effect size (SD units)	-0.12	-0.02	0.02
SAT verbal scores			
Impact (SAT score)	-23.9	-2.7	-3.2
RMSPE	14.1	10.1	12.8
Fit-adjusted *p*-value	0.08	0.33	0.58
N (states)	50	47	47
Effect size (SD units)	-0.16	-0.02	-0.02

Source: Authors' estimates using state-level SAT data.

FPL = federal poverty line; RMSPE = root mean squared prediction error; SD = standard deviation.

Appendix B: Analytic Sample Details

Table B.1. Panel dimensions and modeling prediction error

	States in Sample	RMSPE
10-cohort Panel		
Math	35	2.86
Reading	37	3.35
Average	**36**	**3.11**
11-cohort Panel		
Math	31	2.45
Reading	29	4.21
Average	**30**	**3.33**
12-cohort Panel		
Math	24	4.13
Reading	26	5.08
Average	**25**	**4.61**

Source: Authors' estimates using National Assessment of Educational Progress (NAEP) data obtained from the Urban Institute's "America's Gradebook" NAEP dashboard.

RMSPE = root mean squared prediction error.

Table B.2. Number of states included in state-level data

	States in Full Sample	States in Restricted Sample
Math		
Grade 4	35	33
Grade 8	35	33
Gains	**30**	**30**
Reading		
Grade 4	38	35
Grade 8	36	34
Gains	**28**	**27**

Source: Authors' estimates using National Assessment of Educational Progress (NAEP) data obtained from the Urban Institute's "America's Gradebook" NAEP dashboard.

Table B.3. Number of counties included in county-level data

	Counties in Full Sample	Counties in Restricted Sample
Math		
Grade 4	60	50
Grade 8	40	40
Gains	**30**	**30**
Reading		
Grade 4	60	60
Grade 8	50	40
Gains	**50**	**40**

Source: Authors' estimates using restricted-use student-level National Assessment of Educational Progress (NAEP) data obtained directly from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1992–2017 Mathematics and 1994–2017 Reading Assessments.

Notes: Numbers rounded to the nearest 10 to protect confidentiality.

Figure B.1. NAEP state-level reading data: 10 cohorts of students in grades 4 through 8

	Grade levels by year and NAEP cohort										
Cohort	**'94**	**'98**	**'02**	**'03**	**'05**	**'07**	**'09**	**'11**	**'13**	**'15**	**'17**
1	**4**	**8**									
2		**4***	**8**								
3		**4***		**8**							
4			**4†**		**8**						
5				**4**		**8**					
6					**4**		**8**				
7						**4**		**8**			
8							**4**		**8**		
9								**4**		**8**	
10									**4**		**8**

Note: Bold vertical line indicates implementation of reforms in 2007.

* Grade 4 scores in 1998 used for both 2002 and 2003 grade 8 cohorts, representing four- and five-year gaps, respectively.

† Grade 4 scores in 2002 used for 2005 grade 8 cohort, representing a three-year gap.

Mathematica

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom

mathematica.org