School Autonomy in Denver: The Impact of Innovation Schools

Working Paper 240

National Center for the Study of Privatization in Education

Teachers College, Columbia University

June 10, 2020

Philip Gigliotti

Rockefeller College of Public Affairs and Policy

University at Albany, SUNY

Abstract

Improving performance in struggling urban schools is one of the most persistent challenges in education. In Denver, Colorado, a program called Innovation Schools provided urban public schools with autonomy to implement comprehensive managerial and educational reform plans by waiving district policies. I evaluate the Innovation Schools reform using a difference-in-differences design and find that the program increased end-of-year standardized test scores by 0.1 to 0.3 standard deviations in Innovation Schools. However, exploration of how these effects developed over time suggests that the early impacts of the program faded out following year two of implementation. These findings suggest the program was able to rapidly turn around low-performing schools, but these schools struggled to sustain results. This suggests school turnaround may be possible, but schools need to sustain efforts over time to preserve results.

JEL Codes: I22, C23.

Acknowledgements: I would like to thank Erika Martin, Alan Wagner, Beth Schueler, Michael Kang, and Andre Kiesel for their comments on this draft.

Suggested citation: Philip Gigliotti, School Autonomy in Denver: The Impact of Innovation Schools, Working Paper 240, National Center for the Study of Privatization in Education, Teachers College, Columbia University, June 10, 2020, accessed at http://ncspe.tc.columbia.edu/working-papers/WP240.pdf.

Introduction

Low performance in urban public schools is a persistent educational problem, driven by persistent achievement gaps between wealthy and poor students. Recent studies investigate whether it is possible to "turn around" these schools, by implementing comprehensive managerial interventions to rapidly improve performance (Heissel and Ladd 2018, Schueler, Goodman, and Deming 2017, Zimmer, Henry, and Kho 2017, Strunk et al. 2016). This study evaluates a turnaround program in Denver Public Schools (DPS) which provided schools with managerial autonomy to lead innovative reforms. The aim of this study is to shed light on the impact of managerial autonomy on public school performance and the possibility of leading successful turnaround reforms. The possibility of turnarounds to improve performance in struggling urban schools would represent progress after decades of unsuccessful efforts.

Many interventions have attempted to improve the performance of struggling urban schools with few encouraging results. States have increased educational aid to high-need schools by thousands in per-pupil dollars with little resulting improvement in national performance trends (Hanushek 1997, Yinger 2004). While LaFortune, Rothstein, and Schanzenbach (2018) now demonstrate that these investments produced achievement gains, their magnitude does not suggest that finance reforms could close achievement gaps without unprecedented investments. Other reforms focus on holding schools accountable for their performance on standardized examinations. While these accountability reforms have been shown to increase performance in certain contexts (Chiang 2009, Rockoff and Turner 2010), evaluations of two national accountability efforts exhibited failure to close achievement gaps or ambiguous impact on student achievement. (Dee and Jacob 2011, Hanushek and Raymond 2005). Furthermore, even after achieving universal implementation of accountability systems in American public schools following the 2001 No Child Left Behind Act (NCLB), achievement gaps still persist.

Some reformers advocate for charter schools, public schools which make decisions independently of school districts, allowing them to implement programs which are responsive to students' needs. While they are subject to criticism, and often demonstrate negative or null impact on student achievement, (Bifulco and Ladd 2006, Imberman 2011, Booker, Gilpatric, and Gronberg 2007) many studies now demonstrate that charter schools can improve student outcomes in a variety of contexts (Dobbie and Fryer 2011, Angrist et al. 2016, Booker et al. 2011).The potential for charter schools to improve student outcomes may work through two mechanisms. The first is school choice: by allowing students to enroll in good schools and leave bad schools, competition for students may drive schools to improve performance. The second is independence from centralized district management, which allows charters to be responsive to student needs, making them more effectively and innovatively managed (Chubb and Moe 1990).

Empirical evidence suggests that innovative management practices can sometimes explain the positive impacts of effective charter schools. Dobbie and Fryer (2011) have linked the positive effects of high-quality charter schools to a set of innovative managerial strategies. In a project called Apollo 20, involving 9 district high schools and 11 district elementary schools in Houston, Fryer (2014) demonstrated that these strategies could be adopted by traditional public schools and produce similar positive effects: students posted significant annual gains in math but little progress in reading. Without the buy-in of parents and students opting into a charter school, Apollo 20 necessarily exhibited limitations (Abrams, 2016). The project nevertheless suggests that providing traditional public schools with the autonomy to implement managerial reforms could play an important role in turnaround efforts and lead to performance improvements

In 2009, DPS implemented a program called Innovation Schools that allowed traditional public schools to waive district policies in order to pursue innovative managerial reforms, such as comprehensive changes to curriculum, scheduling and personnel management. This reform contributed to district-wide efforts to improve performance in struggling schools through enhanced accountability and innovative school governance, including expanded school choice. Innovation schools were accountable to DPS for faithful implementation of comprehensive reform plans and improved performance according to a district-wide accountability system which gave each school an annual performance rating and set targets for growth. The conceptual model of change underlying the Innovation Schools program was that providing traditional public schools with the autonomy to pursue independent managerial reforms would lead to improved school performance in the context of a struggling urban school district.

Evaluation of the Innovation Schools program provides multiple contributions to the educational literature. First, it contributes to the literature on school turnarounds. School turnarounds are interventions that attempt to rapidly improve the performance of struggling public schools using comprehensive managerial reforms. The Innovation Schools program can be considered a turnaround-style reform, as it leverages managerial reform to rapidly improve school performance in the context of a very low-performing district. Estimating the impact of this program, including how program impacts develop over time, can expand the turnaround literature and suggest new strategies for turnaround reforms.

School turnarounds have been evaluated in a number of contexts with mixed results. The most notable implementation of school turnaround occurred under the Title 1 School Improvement Grants (SIG) program, which provided grants to persistently low-performing schools which were conditional on managerial reform, often including replacement of staff and leadership. SIGs produced positive academic impacts in Ohio (Carlson and Lavertu 2018) and California (Dee 2012), but evaluation of a national sample of SIG schools showed null academic impacts and indications of failed implementation (Dragoset et al. 2017). Turnarounds have been evaluated in other contexts with some showing large impacts on academic performance (Schueler, Goodman, and Deming 2017) but others showing mixed academic impacts and negative impacts on school climate (Zimmer, Henry, and Kho 2017, Strunk et al. 2016, Heissel and Ladd 2018). These mixed results suggest expanding the turnaround literature could refine understandings of the factors leading to successful or unsuccessful reforms.

Evaluation of the Innovation Schools program provides an opportunity to explore the understudied relationship between managerial autonomy and school performance in public education, and how this autonomy interacts with turnaround strategies. Autonomy can be understood as decentralization of control from higher levels of governance, in this case school districts, to lower levels, in this case schools. The operational principle underlying provision of autonomy is that lower levels of governance can use discretion to implement policies that are more responsive to conditions at the point of service delivery. Since the key policy lever in this program is a waiver from district policies, the program allows schools to employ discretion to implement independent managerial reforms that are more responsive to student needs than standard district practices. Managerial autonomy was linked to performance improvements in a study of British public schools (Clark 2009), however studies of public schools in Chicago and Boston showed null effects on academic performance (Abdulkadiroglu et al. 2011, Steinberg 2014). Evaluation of the Innovation Schools reform can assess the efficacy of autonomy-based interventions as a school turnaround strategy.

I evaluate the DPS Innovation Schools intervention, providing estimates of treatment effects on academic outcomes and how these effects develop over time. Using a two-way fixed-effects difference-in-differences design, this study assesses how school performance changes in schools transitioning to Innovation status, a result which assesses the efficacy of the program as a managerial reform to improve struggling schools. Results suggest that standardized test scores in math, reading and writing improved by 0.1 to 0.3 standard deviations (sd) in public schools that transitioned to Innovation status, indicating the program rapidly improved performance. Effects greater than 0.2 sd are considered large relative to other studies in the educational literature, according to the framework of Kraft (2020). However, exploration of how these effects develop over time suggests program effects peaked during year 2 of implementation and declined steeply in following years. The finding that turnaround effects are susceptible to fade-out is a new insight that can inform future turnaround evaluations, and is a major contribution of this study. This study provides evidence on the effects of managerial autonomy on public school performance and the potential for turnaround reforms to drive improvements in struggling urban schools.

Denver Innovation Schools

In 2008, the State of Colorado passed the Innovation Schools Act, which allowed public schools to submit formal plans to improve student achievement through innovative managerial reforms. Upon approval, Innovation Schools were provided waivers exempting them from district policies, allowing greater autonomy over staffing and other operational practices. While the Innovation Schools Act allowed all Colorado school districts to create programs, Denver Public Schools (DPS) was the first to implement it and authorized more Innovation Schools than any other district. While 3 other districts implemented programs during the period of this study, only 1 had a significant number of Innovation schools, models were inconsistent between districts, and the majority of schools opened late in the period of this study (during the 2012-13 academic year). For these reasons, this study focuses only on the Denver reform.

The Innovation Schools program was part of a broader DPS strategy using school choice and accountability to drive performance improvements. These reforms were implemented to address a performance crisis in the district, which drove enrollment declines that threatened the sustainability of district programming. The Denver reforms reversed enrollment loss, with enrollment growing from about 80,000 to 90,000 during the period of this study spanning the 2006-07 to 2013-14 academic years. Evaluation of Denver's reform model suggests that school choice initiatives associated with the reform improved achievement and shows that the DPS context and portfolio system offers enough variation in governance regimes and a large enough sample size for analysis as a self-contained ecosystem (Abdulkadiroglu et al. 2017).

Denver Innovation Schools are public schools that are provided exemptions from district policies in order to pursue innovative managerial reforms. While Innovation Schools are one of many school models available to students in the DPS, they are not charter schools. Some Innovation Schools were traditional public schools that petitioned a change to Innovation status, while others were new schools that opened as Innovation Schools. By 2014, approximately 30 schools had attained Innovation status out of approximately 200 in DPS. Figure 1 charts the openings and transitions of these schools. Innovation Schools are distinct from charter schools in their level of managerial autonomy. While Innovation Schools attain some autonomies granted charter schools under Colorado law, districts maintain oversight over the schools and their leadership. Schools receive this status by submitting a comprehensive reform plan subject to district approval. All traditional public schools in DPS are eligible to submit Innovation Plans, and new schools may submit plans to open with the enhanced autonomy offered under the program. While denial of a proposal is possible, there is no documentation of denials during the period of the study. Once the district approves an Innovation Plan, schools are subject to performance review and renewal every 3 years. During the period of this study, no Innovation Schools dropped out of the program. (Colorado Department of Education 2015)

Under Innovation Plans, schools request waivers from district policies. The most common waivers are for policies related to human resources practices, school day and year length, and curriculum requirements. Schools may not waive policies related to accountability requirements, student safety, or teacher retirement. Table 1 lists the types of waivers received and percentage of DPS Innovation Schools receiving them (Colorado Department of Education 2013). Innovation Schools receive waivers from policies relating to budgeting, including teacher collective bargaining, which they use to adjust teacher compensation, allocate resources for critical hires, compensate teachers for added instructional time, and contract out for services like food, maintenance and security. The powers to waive teacher collective bargaining were granted under the Innovation Schools Act. Schools receive instructional waivers which they use to implement innovative curricula including project-based and student-centered learning, augment instruction with educational technologies, and offer electives and enrichment such as concurrent enrollment programs. Innovation schools use human resource management waivers to create their own teacher evaluation systems, pay for performance schemes, and to protect their teachers from being reassigned to other schools. All Innovation Schools sought scheduling waivers to allow them to extend the school day and school year, most commonly to provide an extra 30 minutes per day of instruction in math and ELA. Finally, some schools received waivers from restrictions on receiving non-tax revenue, which allowed them to create foundations and engage in external fundraising (Colorado Department of Education 2010).

Some used these extra funds to engage outside consultants to help with strategy and program development. One example was the Denver Summit Schools Network (DSSN), which contracted with the Blueprint Schools Network (BPSN) (https://blueprintschools.org), a private nonprofit educational management organization. BPSN worked in partnership with Harvard University's EdLabs to lead turnaround reforms in struggling public schools according management principles informed by the Fryer (2014) research agenda. The DSSN was initiated in the 2011-12 academic year with 11 schools (2 of which transitioned from traditional public schools to Innovation status and are included in the treatment group of this study) and operated for the duration of the study period. Fryer (2014) evaluated the impacts of Denver DSSN schools, some of which were Innovation Schools, and found positive impacts.

DSSN schools paid $800,000 per year to the BPSN, funded in part by $6.7 million in School Improvement Grant (SIG) funding and $4.2 million in private fundraising (none of the SIG funded DSSN schools are included in the treatment group of this study, and only 1 treatment school was SIG-funded (Trevista at Horace Mann)). DSSN schools benefitted from 75 full time math tutors, students began the school year 6 days early and spent an extra hour in class, and most of the staff and teachers in the schools were replaced. (Tomassini 2012, Robles 2011) To understand the impacts of different types of Innovation Schools and make a contribution beyond Fryer (2014), I estimate separate models that assess whether there were heterogenous effects for DSSN- and SIG-funded schools.

In addition to district oversight of Innovation Plans, Innovation Schools are subject to the Denver-specific accountability system called the School Performance Framework (SPF). The

SPF rates schools annually based on standardized testing performance. In 2014, 10 Innovation schools were rated at the lowest performance rating and only 1 graduated from a turnaround rating in the prior period. However, most of the Innovation Schools in Turnaround Status were schools that had opened as Innovation Schools. Of the 13 schools that transitioned from traditional public school to Innovation status, only 2 were rated turnaround, and some showed evidence of improvement. The experience of schools transitioning to Innovation Status is most relevant to understanding the success of the program as a managerial intervention to turn around struggling schools. Improving performance under SPF factors heavily into evaluation and renewal of Innovation Plans every 3 years (Colorado Department of Education 2015).

The reforms in DPS mirror a national trend towards portfolio management (PM) reforms, which have been implemented in cities including Los Angeles, Chicago, Indianapolis, and New Orleans. In these reforms, school districts allow low performing schools varying levels of autonomy over staffing, curriculum, budgeting and operational decision-making. This autonomy is generally coupled with accountability requirements: if schools fail to improve performance, their autonomous status may be revoked (Marsh, Strunk, and Bush 2013). PM reforms were evaluated in Los Angeles with disappointing achievement results (Strunk et al. 2016). However, PM reforms in New Orleans, where the district was almost entirely converted to charter schools with significant managerial autonomy, demonstrated performance improvements (Harris and Larsen 2016). PM reforms are closely related to school turnarounds, which leverage managerial reforms to drive rapid improvement in struggling schools. Turnarounds have shown evidence of success in a number of instances (Carlson and Lavertu 2018, Schueler, Goodman, and Deming 2017, Dee 2012), though the results in sum have been mixed (Heissel and Ladd 2018, Dragoset et al. 2017, Zimmer, Henry, and Kho 2017).

Fryer (2014)'s evaluation of DSSN schools, which included Innovation Schools, found significant effects in math and null effects in reading. Abdulkadiroglu et al. (2017) evaluated the DPS context in a study of charter school effects, which included a supplementary model isolating effects of Innovation Schools that found null negative and null positive results depending on specification. This study will evaluate 13 schools that transitioned from traditional public schools to Innovation Schools between the 2006-07 and 2013-14 academic years. This variation in Innovation School status within schools allows estimation of a causal effect using a two-way fixed-effects difference-in-differences design. These estimates differ from the estimates of Fryer (2014) and Abdulkadiroglu et al. (2017). This study includes a longer panel than either study, with both a longer pre-treatment and post-treatment period. The analytic sample only includes 2 of 7 Innovation Schools evaluated by Fryer, whose study leveraged student-level rather than school-level variation in Innovation Schools status and could therefore analyze schools that opened as Innovation Schools and experienced no variation in treatment. It provides new estimates for 11 schools not evaluated by Fryer, all of which experienced transition to Innovation status, and whose experience can suggest the ability of the program improve performance in struggling schools, consistent with a turnaround model. The study uses a different identification strategy than Abdulkadiroglu and colleagues, who used lottery randomization to estimate student-level achievement effects. It also uses a different sample; lottery randomization allowed them to include Innovation Schools without variation in treatment status in their estimates. Their counterfactual indicates how a student's performance would differ if he or she attended another district school rather than an Innovation School, while my counterfactual indicates whether Innovation Schools improved from their pre-treatment performance level. It is simultaneously possible for Innovation Schools to improve following treatment, but fail to reach parity with other district schools, which could lead to a negative estimate in the Abdulkadiroglu model. Thus, my identification strategy is most relevant to understanding how the program changes performance within schools over time, rather than how attending an Innovation School impacts student achievement compared to other schools.

Since my analysis examines schools that transitioned from a traditional public school to Innovation status, my results reveal whether the Innovation Schools program can turn around struggling public schools, rather than how the performance of students entering an Innovation School changes. My analysis probes the sensitivity of Innovation School effects to a greater extent than prior studies, which addressed the program briefly in supplementary analyses. I pay careful attention to parallel trends assumptions, non-random selection into treatment, treatment heterogeneity and inferential difficulties arising from the small number of schools experiencing change in treatment status. I demonstrate that managerial autonomy leading to innovative reforms may be an effective strategy to turn around low-performing urban schools, but that performance improvements may deteriorate without continuous effort to sustain results.

Data

This study uses data collected alongside annual accountability requirements by the Colorado Department of Education (CDE), which are publicly available on their data website (www.cde.state.co.us/cdereval). I extract data based on Enrollment, Demographics, Resources, Staffing and Achievement from the 2006-07 to 2013-14 Academic Years, resulting in a final sample of 148 schools in DPS, including all schools which transitioned to Innovation status.

The study estimates the effect of Innovation Schools treatment on academic achievement, using test scores on the CSAP (Colorado Student Assessment) and TCAP (Transitional Colorado Assessment Program) exams, Colorado's year end standardized tests in writing, reading and math, taken by students in grades 3-10. The CSAP changed to TCAP in the 2011-12 academic year, but the scale and distribution of test scores changed very little and there is little evidence of differential changes pertaining to treated and untreated schools. These measures were available as school level means through the CDE's data portal (Colorado Department of Education n.d.). As school-level averages of an exam with uniform scale between grades, these measures are comparable across all schools including elementary, middle, and high schools. To remove temporal and distributional trends, I standardize these measures with respect to full sample (all DPS schools) by year consistent with prevailing norms for standardization of test scores in the educational policy literature. I also test for effects on non-academic and organizational outcomes using the following dependent variables: enrollment, student-teacher ratio, teacher salary (adjusted to 2016 dollars) and discipline (a measure of combined suspensions and expulsions during the academic year). I am missing data from the 2006-07 academic year for some of these variables. The organizational models, accordingly have one less pre-treatment year.

The study uses a binary indicator of treatment status coded 1 following approval of an Innovation Plan and 0 otherwise. This is equivalent to an interaction between a binary indicator of membership in the Innovation Schools treatment group, and binary indicator for each school's post-treatment period. While there were 30 Innovation Schools in DPS during the period of the study, only 13 demonstrate variation in treatment status indicating transition from a traditional public school to an Innovation school (4 elementary Schools, 5 schools with elementary and middle grades, 1 middle school, 1 school with middle and high school grades, and 2 high schools). The other 17 schools were new schools that opened as Innovation Schools, and thus do not have a pre-treatment period. Since fixed-effects models use only within-panel variation, Innovation Schools without treatment variation will not contribute to identification and will enter into the estimate as equivalent to untreated schools. To preserve homogeneity of treated and comparison groups, I exclude all Innovation Schools with no variation in treatment status from the analysis. The results with this specification are fundamentally equivalent to models that include Innovation Schools with time-invariant treatment status. The final treatment group sample includes all 13 DPS schools that transitioned from a traditional public school to an Innovation School during the period of the study, and therefore provide within-school variation in treatment status to leverage in a two-way fixed-effects design.

Due to limitations of publicly available data, I have few control variables to include in my models. However, I can include enrollment, student teacher ratio, % free lunch eligibility, % black students and % Hispanic students. Given model assumptions hold, lack of rich controls does not prevent identification, though I will be unable to tell how the program impacted the proportion of students with disabilities or English language learners and how this contributed to treatment effects. This is a limitation of my analysis. However, I can test for changes in composition based on % free lunch, % black and % Hispanic students (Appendix Table 1) and find no evidence of changes in student composition.

Methods

Ordinary least squares estimates of Innovation School impacts on school performance may be biased since participation in the program was not randomly assigned and may therefore be correlated with unobserved characteristics of participating schools. To address the endogenous nature of the reform, I specify a two-way fixed-effects difference-in-differences model according to the following specification:

$$y_{st}=\gamma_0 Innovation\ School*Post_{st}+\gamma_1 X_{st}+\theta_s+\tau_t+\varepsilon_{st} \qquad \text{(Equation 1)}$$

In this equation, y_{st} is an outcome of interest for school s in year t, X_{st} is a vector of school level demographic and organizational characteristics, and ε_{st} is a stochastic error term for school s in year t. The identification strategy leverages within-school transition from a traditional public school to an Innovation School. The measure $Innovation\ School*Post_{st}$ equals one following transition to Innovation School status and zero prior for each treated school, and zero for all untreated schools in all periods. Since transition occurs at different times for different Innovation Schools, this indicator is equivalent to an interaction between a time invariant indicator of treatment group membership and each school's unique post-treatment period. θ_s is a vector of school fixed-effects which absorb all time invariant characteristics of each school, including the binary indicator of treatment group membership included in traditional difference-in-difference models. τ_t is a vector of year fixed-effects, which absorb temporal factors shared across schools, including the binary indicators for all post-treatment periods for all treatment cohorts included in traditional difference-in-differences models. Since the school fixed-effects absorb time invariant treatment status and the year fixed-effects absorb all possible post treatment periods, γ_0 can be interpreted as a difference-in-differences estimator. This functional form attempts to mitigate selection into Innovation Schools treatment to capture treatment effects with reduced bias, and allows for estimation in the presence of staggered treatment initiation.

The identification assumptions of this model are that treatment is assigned exogenously conditional on school and year fixed-effects and that treated and untreated schools would have similar outcome trends in the absence of treatment. I assess the plausibility of these assumptions by showing robustness of effects to different comparison groups, by testing for parallel trends violations in the pre-treatment period with placebo tests, event-study regressions and graphical analyses, and by probing robustness to school-specific linear time trends and lagged dependent variables. For the placebo tests I estimate treatment effect models with "placebo" treatment dummies assigned to the treatment group in each pre-treatment year when no actual intervention existed. These dummies are equivalent to an interaction between a time-invariant treatment group indicator and a full vector of year fixed-effects, consistent with the following specification (All terms equivalent to Equation 1, $\gamma 0$ is a vector of placebo coefficients):

$$yst=\gamma 0 Innovation\ School * Year=2007-08 Year=2012-13 Yearst+\ \gamma 1 Xst+\theta s+\tau t+\ \varepsilon st$$

(Equation 2)

Treated schools are removed from the sample as they enter a genuine treatment period; the panel terminates in the 2012-13 academic year. Estimating a significant effect in a placebo period, where no treatment existed, could indicate a parallel trend violation (Mora and Reggio 2017).

I also include event-study regressions by estimating two-way fixed-effects models with treatment dummies indicating each treatment school's year relative to treatment, ranging from 7 years pre-treatment to 5 years post treatment. The treatment group for these models is equivalent to the main models, consisting of schools who transitioned to Innovation School status. These indicators are coded as one for each Innovation School in its nth year relative to treatment, and zero for Innovation schools not in their nth year relative to treatment or untreated schools. This is equivalent to an interaction between the time invariant treatment group indicator and a binary indicator of each school's year relative to treatment, consistent with the following specification (All terms equivalent to Equation 1, $\gamma 0$ is a vector of event-study coefficients):

$$y_{st}=\gamma_0 \, Innovation\ School * \sum_{i=-7,\ i\neq -1}^{5} Treatment\ Year_{ist} + \gamma_1 X_{st}+\theta_s+\tau_t+\varepsilon_{st}$$

(Equation 3)

Innovation Schools in the first treatment cohort (treated in the 2009-2010 academic year) would be coded as one for 3 pre-treatment indicators and all 5 post-treatment indicators, given an 8-year panel spanning the 2006-07 to 2013-14 academic years. Innovation schools in the final treatment cohort (treated in the 2013-14 academic year) would be coded as one for 7 pre-treatment indicators and 1 post-treatment indicator. Because this requires excluding a year as the reference category, I estimate two different specifications. The primary model excludes the last pre-treatment year (year negative 1) excluded. For robustness, I estimate a second model with the third pre-treatment year excluded (year negative 3), since this is the earliest pre-treatment year shared by every treated school. In the absence of pre-treatment trend violations, we should expect to see null coefficients in the pre-treatment years, especially close to the treatment window, and treatment effects in the post treatment year similar to those in the main models.

I also estimate models with school-specific linear time trends. Introduction of a panel-specific trend in difference-in-difference models allows for linear violations of the parallel trends assumption; schools can be moving in opposing paths and identification assumptions are maintained, provided there are no quadratic deviations (a more plausible assumption than no linear deviation) (Mora and Reggio 2017). Finally, I estimate models with a lagged dependent variable, which accounts for prior year achievement in the estimates and may provide more robust treatment effect estimates in the presence of parallel trends violations. With a relatively long panel, and an intention to probe robustness to prior achievement rather than estimate causal effects, concerns about Nickell bias likely do not warrant a dynamic panel approach to these supplementary analyses. These models are provided in a robustness checks section.

Since schools are not treated randomly, but rather adopt Innovation Schools treatment by choice, and because there are significant differences between treated and untreated schools on observable characteristics, I estimate my models in 2 different subsamples. The first subsample compares Innovation Schools to all other schools. The second subsample is chosen using propensity score matching. I estimate propensity scores using a probit model with all covariates and dependent variables employed in the main models, estimating the propensity scores separately for each treatment cohort starting with the first, restricting the sample to the last pre-treatment year for each treatment cohort, and excluding matched schools from prior cohorts from the potential matches of future cohorts to avoid excessive overlap. I select the 5 nearest neighbor matches with replacement. I then absorb the entire panel of any school that is a nearest neighbor match for any treated school according to this strategy. This yields a matched sample of 62 clusters with the comparison group including the five nearest propensity score matches to each treated school, with some control schools serving as matches to multiple treated schools. I also checked robustness to 2 other comparison groups selected on the basis of free lunch eligibility and test scores, and the results were fundamentally equivalent (results available on request).

The literature on school turnarounds shows that effects often grow in the years following implementation (Carlson and Lavertu 2018, Sun, Penner, and Loeb 2017). Alternatively, if turnaround schools decrease effort following an initial implementation period, treatment effects may decrease as time progresses. The event-study models described above allow heterogeneity of treatment effects by number of years since treatment initiation. The estimates on the indicators for the first through fifth years of treatment allow interpretation of the cumulative effects of the program in each year since treatment initiation. I also explore whether the effects of the program were larger or smaller in certain schools. To explore this heterogeneity, I engaged in list-wise deletion of schools from the treatment group. If removing the school decreased either the effect size or significance level of the main math effect, indicating that the school's performance was contributing to the positive average treatment effect, I added the school to my "High Performer" group. If removing the school increased both the effect size and significance level, I added it to my "Low Performer" group. This allows for interpretation of whether effects of the program were evenly distributed, or whether they masked both successes and failures. Since many of the Innovation Schools were externally managed by the Blueprint Schools Network, a CMO-like private nonprofit management group, it is important to understand how effects of the program are driven by these externally managed schools. To assess this dynamic, I estimate separate treatment effects for Blue Print and non-Blue Print schools. I include these three results in a section on treatment heterogeneity following the main results.

All models are estimated with Huber-White robust standard errors clustered by school, to address heteroscedasticity and autocorrelation within schools. However, this has the potential to be inaccurate given the small number of treated clusters in my sample. Significant attention has focused on inference in difference-in-differences models (Bertrand, Duflo, and Mullainathan 2004), particularly when the number of clusters or treated clusters are small (Mackinnon and Webb 2018, Mackinnon 2016, Cameron, Gelbach, and Miller 2008). With a small proportion of treated clusters, cluster robust standard errors can lead to over-rejection in hypothesis testing. Wild bootstrap clustered standard errors can provide more accurate estimates of the standard error with a small number of treated clusters, but can also lead to under-rejection. For this reason, I conduct hypothesis testing on the coefficients of interest using wild bootstrap clustered standard errors and provide the associated p-values corresponding to each estimate. Providing both cluster robust and wild bootstrap hypothesis testing should provide an upper and lower bound for the level of significance of my estimates.

Results

Descriptive Statistics

Summary statistics are provided in Table 2. DPS is a low-performing urban school district, which does not exceed 50% proficiency in writing, reading or mathematics. Students in the district are overwhelmingly poor and minority, with average free or reduced-price lunch eligibility of 71%, average Hispanic composition of 58% and average Black composition of 14%. Schools have 533 students on average; the average student teacher ratio is 17.45, which was above the national average of 16.1 in 2014. DPS has approximately 90,000 students, making them the thirty-fourth largest public-school district in the nation in 2013. Teachers make approximately $53,000 per year, on par with the national average. During the period of the study, DPS spent approximately $6,000 per pupil and Colorado was in the bottom 25% of states by per pupil expenditures. Table 3 presents summary statistics comparing Innovation Schools in the treatment group (those who transitioned to Innovation Status) and non-Innovation schools in the 3-year pretreatment period from the 2006-07 to 2008-09 academic years. Innovation Schools are poorer, more diverse and lower-achieving that non-treated schools. Table 3 also presents comparisons of Innovation Schools to the propensity matched sample used in the analysis. Innovation Schools are almost perfectly matched to comparison schools in this sample on organizational and demographic characteristics. They do have more black and fewer Hispanic students, though only the difference in % black students is statistically significant at the .10 level.

Main Results

In the following section I estimate difference-in-differences models with school and year fixed-effects, and controls for enrollment, student-teacher ratio, % free lunch eligibility, % black students and % Hispanic students with both cluster robust standard errors and Wild bootstrap clustered standard errors (results are robust to estimation without controls, and these estimates are available upon request). Table 4 presents the main results, showing treatment coefficients for writing, reading and math from left to right, with effects shown for each dependent variable in the full sample and propensity score matched sample. All subsequent treatment effect tables will follow this general format. Effects in writing are statistically insignificant but meaningful at about .1 standard deviations (sd) in test score gains. Effects in reading are significant at the 0.05 level in the full sample and at the 0.10 level in the PSM sample (and at the 0.10 level with wild bootstrap hypothesis testing), and about 0.15 sd in magnitude. Effects in math are significant and measure about 0.25 standard deviations. All are significant at the 0.05 level with cluster robust standard errors, and at the 0.05 level with wild bootstrap clustered errors (p= 0.043, 0.041). According to guidelines established by Kraft (2020), effect sizes of educational interventions greater than 0.2 standard deviations are considered large, and fall above the 80th percentile of empirical effect sizes in a review of over 200 studies. As can be seen in systematic reviews of the evaluation literature of educational accountability studies, it is common for educational interventions to show larger results in math than in reading and writing, possibly because math skills are developed in the classroom to a greater extent (Figlio and Loeb 2011).

Treatment Heterogeneity

Table 5 estimates event-study models which reveal treatment effect heterogeneity by years since treatment initiation. (Model coefficients are plotted in figures 5-7. Since coefficients are roughly equivalent in both estimation samples, I only plot the full sample models. Appendix Table 3 and figures 8-10 show the same results with an alternate reference category (earliest shared pre-treatment year excluded) with equivalent results.) These models show large effects developing into year two of program implementation, but indicate trouble sustaining results. The math results show positive effects of approximately 0.15 sd in the first year of treatment, which grow to approximately 0.3 sd in the second and third years. Effects in the fourth year remain large, though marginally significant, but effects regress to approximately zero in the 5^{th} year. In writing, effects increase to statistical significance in the second year of treatment, measuring approximately 0.2 sd, but then recede dramatically in the following years, fully regressing to zero. Effects in reading grow to significance in the second year as well, measuring approximately 0.15 sd, and then follow a similar receding pattern. These results suggest that the effects of Innovation Schools treatment grow over time, especially leading into the second and third years of implementation, but raise concerns about fade-out suggesting difficulty sustaining the success of the program. The treatment group in these models is not stable across post treatment years; for example, while all treatment schools have a first post-treatment year, only the first treated cohort has a fifth post-treatment year due to the staggered timing of treatment initiation. Results by treatment cohort (available on request) show a similar fade-out pattern across treatment cohorts, though the second treatment cohort showed more success at sustaining results.

I also perform exploratory analysis of heterogeneity in the size of treatment effects. The listwise deletion procedure outlined in the methods section left me with a treatment group of 8 "High Performer" schools that showed dramatic improvements under the Innovation Schools program (see Table 6). These schools were Cole Arts and Sciences Academy, Godsman Elementary School, Green Valley Elementary School, McGlone Elementary School, Montclair

School of Academics and Enrichment, Trevista at Horace Mann and Valdez Elementary School and Whitter K-8 school. The effects on these schools were large and highly significant ranging from 0.25 sd in writing to 0.4 sd in math, all significant at the 0.01 level. Most of these schools are elementary schools and middle schools, with younger students who may be more receptive to interventions. I estimate treatment effects for the remaining group, labeled "Low Performers." This group of 5 schools showed negative effects in all subjects, some of which are statistically significant at the 0.10 level. This group includes Centennial (A School for Expeditionary Learning), Grant Beacon Middle School, Manual High School, Summit Academy, and Martin Luther King Jr. Early College. Centennial only had 1 year of treatment, so their poor performance could have been a result of transitional disruptions. Most of these schools include high school grades, where students might be more resistant to interventions or more likely to be impacted by disruptions. These models do not exclude performance declines driven by other factors, and these schools may have had similar declines in the absence of Innovation Schools treatment. These exploratory analyses should be interpreted as descriptive, as its uncertain whether parallel trends assumptions hold for these restricted performance groups.

Three of 8 high performers belonged to the DSSN (Green Valley Elementary School and McGlone Elementary School) or were SIG-funded (Trevista at Horace Mann). This group was selected because every school contributed to the positive effects of the program, but it is important to assess the dependency of the results on these externally supported schools. Main results are not robust to exclusion of DSSN and SIG schools, but are also not robust to excluding the 5 high-performing district schools and including only externally supported schools (results available upon request). Table 7 modifies the "high-performer" group to exclude DSSN and SIG schools to assess changes in effect sizes and significance. Exclusion of DSSN schools reduces effect size by 0.05 to 0.1 sd in all subjects, but results remain highly significant and 0.2 sd or larger. Further exclusion of the SIG funded school leads to similar results, though effects in Reading are larger than in the full sample. These results suggest that DSSN schools may have done slightly better than other high-performing Innovation Schools, and the SIG funded school did about the same or slightly worse. This suggests that external funding or management does not explain the effects of the program in high-performing Innovation Schools.

Robustness Checks

To probe the robustness of these findings, I first interrogate the parallel trends assumption. First, I investigate this condition graphically. In figures 2-4, I present graphs of writing, reading and math scores over time in both comparison and treatment groups and in the multiple samples. I include a vertical line in 2010 which is the first year of Innovation Schools treatment. The graphs show some common tendencies across all variable. The Innovation Schools cohort seems to have experienced a large increase in performance in the first year of the study, 2006-07, which is 3 years pre-treatment. After that, the cohort levels out and achieves a generally common trend with the comparison group, until 2009-10 when a sizeable treatment effect appears to develop. Then, in the later years of the panel, the visible treatment effect appears to recede, consistent with the findings of the event-study models. The finding that the results of the treatment effect models can be clearly visualized in the graphical plots lends confidence in the main findings.

Some concerns remain about pre-treatment spikes and irregularities in the test score plots, though it should be emphasized that these are unadjusted trend lines, and not causal models. Upon closer investigation, the “performance spike” results in part because one of the treatment schools, Manual High School was closed in the 2006-07 year, and therefore did not have data. Since Manual High School has higher mean test scores than the other Innovation Schools, its addition in 2007-08 increased mean performance, which accounts for much of the spike. Trend plots that exclude Manual High School (available on request) show roughly parallel trends in the pre-treatment period for the writing and math variables, though a less pronounced spike persists in reading. The trend plots show suggestive evidence to support identification assumptions, with a stable or downward trajectory immediately pre-treatment followed by a rapid performance increase post-treatment, but to bolster this analysis, I test statistically for parallel pre-trends in the following section using the placebo testing process outlined in the methods section

The placebo testing procedure used in this paper hinges on the premise that, since placebo treatments are assigned in the pre-treatment period when no actual treatment occurred, their inclusion in the model should result in a null coefficient. Null results on all placebo coefficients suggests support for parallel trends assumptions. The placebo test results are included in Appendix Table 2. These tests perform fairly well, with no placebo coefficient rejected at the 0.05 level in any sample for any variable in the first five years of placebo treatment. In the sixth year we see rejection in all models, but the treated group for this coefficient only includes 1 school and one school-year observation. The success of these tests provides statistical evidence in support of parallel trends assumptions.

Event-study regressions in Table 5 & Appendix Table 3 show a similar trend. The models with year negative one excluded perform well, with only 3 out of 42 coefficients rejected at the 0.05 level and 1 rejected at the 0.10 level, results that could be expected due to random chance. All rejected coefficients are in year negative 7 and negative 6, which only include 1 and 2 treated schools which are more than 5 years from treatment initiation, when pre-trend violations are unlikely to influence results so far in the future. The models with year negative 3 omitted perform somewhat worse, though the rejections are concentrated in years negative 5-7, when the treatment groups only include, 1, 2 and 5 treated schools, and where the schools are far from the treatment window. In the 4 pre-treatment years immediately preceding treatment there are only 2 rejections at the 0.10 level out of 18 coefficients, a result that would be expected due to random chance. Event-study plots in Figures 5-10 show very similar trends, with pre-treatment plots stable in the first 5 pre-treatment year and overlapping with zero virtually everywhere, with treatment effects developing into the second post treatment year and declining thereafter. These tests support the validity of parallel trends assumptions and suggest treatment effects aren't being driven by selection of pre-treatment period in the main model or event-study regressions.

Next, I estimate models with school-specific linear time trends (Appendix Table 4). The results in math yield slightly to this test. However, both models are significant at the 0.10 level, as are estimates in writing. Effect sizes are roughly equivalent to the main models, which suggests that trend violations are not meaningfully biasing the coefficient estimates in the main models. We may interpret the reduced significance as resulting in part from larger standard errors, which are produced by the extra demands that a full vector of school-specific trends (1 per panel) places on the data. Effects may also be less precisely estimated as de-trending the data removes useful variation along with confounding trends. The robustness of the effect sizes suggests evidence of a true performance gain in Innovation Schools, though increased demands on the data cause these estimates to be less precise.

Finally, I estimate models with lagged dependent variables in Appendix Table 5 (since these require lagged data, they have one fewer pre-treatment year). In these models, inclusion of the lagged test scores decreases the writing and reading slightly, though they remain about 0.10

sd. Math results remain significant at the 0.05 level, though they are about 0.05 sd smaller. It is unclear whether any diminishing of effect size or significance is due to inclusion of the lagged dependent variable or shortening of the pre-treatment period, though results are largely consistent with the main models. These findings suggest that the effects in the main models, especially in math, are not being driven by prior year achievement trends.

Non-Academic and Organizational Outcomes

To suggest generalizable mechanisms by which the Innovation Schools reform led to performance improvements, I test the effect of Innovation Schools treatment on a number of organizational outcomes (Table 8). While many of the mechanisms reported under the program, such as expanded instructional time, curriculum reform, teacher evaluation, or managerial autonomy in general are unlikely to be observable, some of the available metrics could yield insights if Innovation Schools grew, decreased class sizes, increased salaries through pay for performance, or increased discipline of students. These results should be interpreted as purely descriptive, and I do not explore the plausibility of parallel trends assumptions for these analyses.

Innovation Schools may have grown following treatment, with an imprecise effect of 17.58 students in the full sample. However, analysis of changes in student composition, measured by % free lunch eligibility, % black students and % Hispanic students (Appendix Table 1) do not show evidence of significant changes, which suggests enrollment growth did not impact treatment effect estimates through changes in student demographics. Student Teacher Ratios show null positive results, suggesting class size reductions were not a mechanism by which Innovation Schools improved performance. Teacher Salary shows a significant negative relationship, which is the best evidence of organizational mechanisms produced in this analysis.

While teacher turnover rates or teacher experience were not available, if lower salaries resulted from turnover leading to replacement with less experienced teachers, these teacher salary effects could indicate large changes in the composition of the teaching staff, consistent with other turnover reforms. Event-study analysis by cohort suggests these salary declines may have been larger in early years of the reform (results available upon request), suggesting performance improvements were associated with early reforms which also faded out over time. Finally, discipline shows null negative results, indicating that at the very least, Innovation Schools did not improve performance through harsh discipline. These organizational analyses are not highly informative in understanding the success of the Innovation Schools. It appears that many causal mechanisms of the program, including the ones listed above, or mechanisms fitting broadly under the umbrella of managerial autonomy, are likely to be unobservable.

Conclusion

For decades, efforts to improve performance in struggling urban schools have fallen short of hopes. Studies of high-quality charters have linked positive achievement effects to innovative managerial practices (Dobbie and Fryer 2011) and experimental analysis by Fryer (2014) find that transplanting managerial innovations from charter schools to traditional public schools can drive performance improvements. In this paper, I investigate a similar managerial reform in DPS. Using a difference-in-differences design, I identify large positive effects of the Innovation Schools reform on math, reading, and writing test scores. However, I also find that program impacts peaked in year two of implementation and then faded out dramatically afterward. This finding suggests turnaround reforms are capable of producing large performance improvements, but that results may be difficult to sustain without continued focus in later years of implementation.

This study contributes a methodologically robust evaluation of the effects of the Innovation Schools program on traditional public schools that transition to Innovation status. I address methodological issues related to parallel trends assumptions, non-random selection into treatment, and inferential difficulties resulting from a small number of treated clusters. Through graphical analyses and placebo testing, I find that the intervention was mostly free of pre-treatment trend violations. Results are generally robust to multiple specification, including with propensity score matched comparison groups, event-study regressions, school-specific linear time trends and lagged dependent variable. By using variation from schools that transition to Innovation Status, I contribute evidence that this reform has the potential to turn around struggling urban schools, and that coupling managerial autonomy and innovation has potential as an effective turnaround strategy. However, this study also provides cautionary new evidence that turnaround impacts may be difficult to sustain, and can be lost in later years of implementation. This finding presents an important consideration for policymakers and school leaders implementing turnaround reforms and should be explored in other contexts.

I find that the program resulted in improvements of 0.1 to 0.3 standard deviations on end of year standardized test scores, effects which are often above the 80^{th} percentile (0.2 sd) of over 200 findings reviewed by Kraft (2020). Where significant, results are generally robust to rigorous hypothesis testing with wild bootstrap clustered standard errors. The calculation of treatment effects in standard deviations (sd) allows comparison to the results of Fryer (2014). While Fryer estimates effects of 0.172 (0.065) sd in math and 0.076 (0.052) sd in reading, I find significant effects of approximately 0.25 sd in math and 0.15 sd in reading, with marginally significant effects of approximately 0.1 sd in writing. These effects are larger than those of Fryer (2014), and achieve significance in reading, where he found null effects. However, evidence from both graphical plots and heterogeneity analyses suggests that the effects of the reform may fade out over time. Changes in teacher salary suggest large shifts in workforce composition in earlier years of the reform which may have also faded over time, suggesting an association between performance increases and organizational changes that were not sustained over time. While turnarounds may be able to rapidly increase performance in struggling schools, sustaining those improvements may require different approaches. Future research on school turnarounds should investigate whether this fade-out is a regular phenomenon in other contexts.

The results contrast with those of Abdulkadiroglu et al. (2017) who found null impacts of Innovation Schools on student achievement. To reconcile these differences, it is important to remember the differences between the two results in terms of sample and identification strategy. Abdulkadiroglu and colleagues employ a sample including both Innovation Schools that transitioned from traditional public schools and a larger group that opened as Innovation Schools. The sample in this study includes only those Innovation Schools that transitioned to Innovation status, which allows estimation of a parameter indicating the efficacy of the program as a managerial intervention to turn around struggling public schools. Established public schools that initiate reforms under the program may use autonomy in different ways or benefit from prior experience leading to better results; brand new schools are likely to suffer from growing pains and experience higher risk of failure than those with an established track record.

Furthermore, the identification strategy in this study leverages transition over time to Innovation status, so results can be interpreted as within-schools improvements from baseline performance. While some researchers may find the lottery randomization used by

Abdulkadirolgu and colleagues more compelling, it estimates a different parameter, which should be interpreted as a comparison between the performance of two identical students during the same time period, one of whom attends an Innovation School and one who does not. If Innovation Schools have lower average performance than comparison schools holding all confounding factors constant, the Abdulkadiroglu method will find a negative impact on student achievement. However, my within-schools difference-in-differences model holds differences in mean performance constant, instead comparing schools on their rate of change from the pre-treatment period. It is possible that Innovation Schools could demonstrate considerable growth from their pre-treatment level over time, even if their ultimate performance mean remains lower than other district schools. Since turn around reforms intend to initiate performance growth in very low-performing schools, it is important to know that this growth is possible, even if it will not lead to parity with other district schools.

These findings suggest the Innovation Schools reform can produce growth in school performance. However, policy makers should be aware that autonomy and innovation involves risk, and the effects of the intervention were heterogenous, driven by a group of successful schools whose progress masked a smaller group with negative impacts from possibly failed reforms. The potential of the program to lead to both successes and failures was born out in anecdotal experiences of Innovation Schools. For instance, Manual High School experienced notable problems when splitting the school into 3 smaller schools hosted in the same building, and its performance declined after transition to Innovation status. (Robles 2011) The proposition of managerial autonomy-based interventions as a high-risk high-reward venture should be understood by policy-makers considering similar programs. The analysis also sheds light on the importance of external support, such as the private Blueprint Schools Network and the School

Improvement Grant program. While these supports possibly contributed to the effects of the program, they do not explain them. Externally supported schools did about the same as other high-performing Innovation Schools. This finding makes a contribution to understandings of privatization in public education and lends new insight into the role of Charter Management Organization-like private actors in turnaround efforts in traditional public schools.

The analysis in this paper is subject to limitations. Most notable is the small sample of treated schools. While the estimated effects are compelling and robust, they are only observed in a small sample of 13 treated schools. The extent to which the experience of 13 schools can be generalized nationwide should be considered with caution. Since similar models have been implemented in other cities, scholars should compare results from those programs to the results seen Denver. Further limitations stem from the limits of publicly available data. I was not able to control for demographic percentages of students with disabilities or limited English proficiency, and I was not able to assess how their populations changed in Innovation Schools. Since schools subject to performance pressures have been documented to manipulate enrollments of these populations, it would have been better to have these variables available (Figlio and Loeb 2011). However, I was able to test for changes in student composition based on poverty and black and Hispanic composition and found no evidence of sorting. Similarly, I am limited in only being able to use high-stakes tests as a measure of performance. Its well-documented that schools can sometimes respond to interventions by increasing performance on high-stakes tests, without demonstrating gains on more general measures of performance (Figlio and Loeb 2011). Future analysis should investigate whether turnaround reforms lead to general performance improvements, or only improvement on high-stakes assessments.

This study contributes to the literature on the impacts of managerial autonomy in public schools. Results suggest a reform involving decentralization of school management from districts to schools produced large performance gains as a turnaround strategy. The relationship between managerial autonomy and public school performance is understudied, with the finding of positive effects in British public schools (Clark 2009) and two findings of null impacts in the American context (Steinberg 2014, Abdulkadiroglu et al. 2011) being the most notable results. This study contributes to what will hopefully be a growing body of literature on this question.

This paper also contributes to literatures on school choice, demonstrating that charter-like managerial interventions can improve student outcomes in struggling urban schools. As studies now consistently find that charter schools can have positive effects on student achievement in multiple contexts (Abdulkadiroglu et al. 2017, Angrist et al. 2016, Dobbie and Fryer 2011, Booker et al. 2011), successful programs should be analyzed for lessons that can improve traditional public schools. Lessons on curriculum, staffing and organizational management can inform efforts to improve performance in struggling public schools. In the case of Innovation Schools in DPS, implementing charter-like management practices such as increased instructional time, student-based learning and customized evaluation schemes can lead to performance improvements. These are generalizable lessons that can be exported to other contexts and brought to scale. While Innovation Schools in the DSSN raised $4.2 million in external fundraising, fundraising may be difficult to sustain over time or may be more difficult in communities with limited resources, as has been experienced by resource-intensive charters such as KIPP as they have expanded operations (Abrams 2016). Furthermore, many of the reforms in Denver would have been impossible without the ability to waive teacher collective bargaining rights, which would be potentially prohibitive in many institutional contexts.

Finally, the study contributes to a growing literature on turnaround reforms (Schueler, Goodman, and Deming 2017, Carlson and Lavertu 2018, Heissel and Ladd 2018). Identifying reforms that can improve performance in struggling urban schools has been an elusive goal in education policy. While there are mixed results, a number of studies now demonstrate that it is possible to rapidly improve public school performance by leveraging comprehensive managerial interventions. This study is limited in its ability to explore mechanisms, and thus cannot provide statistical evidence to distinguish results from other contexts where reforms did not lead to performance improvements. However, the study shows that turnaround is possible in a new context, providing detailed exposition of an autonomy-based model that has demonstrated some efficacy. While effects are heterogenous and fade out over time, and questions remain about the possibility of scaling up from a limited reform, turnarounds may be possible in other settings, especially if further studies can identify determinants of successful reforms.

References

Abdulkadiroglu, Atila, Joshua D. Angrist, Susan M. Dynarski, and Thomas J. Kane. 2011. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." *Quarterly Journal of Economics* 126 (2):699-748.

Abdulkadiroglu, Atila, Joshua D. Angrist, Yusuke Narita, and Parag A. Pathak. 2017. "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation." *Econometrica* 85 (5):1373-1432.

Abrams, Samuel E. 2016. *Education and the Commercial Mindset*: Harvard University Press.

Angrist, Joshua D., Sarah R. Cohodes, Susan M. Dynarski, Parag A. Pathak, and Christopher R. Walters. 2016. "Stand and Deliver: Effects of Boston's Charter Schools on College Preparation, Entry and Choice." *Journal of Labor Economics* 34 (2):275-318.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1):249-275.

Bifulco, Robert, and Helen F. Ladd. 2006. "The Impacts of Charter Schools on Student Achievement: Evidence from North Carolina." *Education Finance and Policy* 1 (1):50-90.

Booker, Kevin, Scott M. Gilpatric, and Timothy Gronberg. 2007. "The Impact of Charter School attendance on Student Performance." *Journal of Public Economics* 91 (5-6):849-876.

Booker, Kevin, Tim R. Sass, Brian Gill, and Ron Zimmer. 2011. "The Effects of Charter High Schools on Educational Attainment." *Journal of Labor Economics* 29 (2):377-415.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90 (3):414-427.

Carlson, Deven E., and Stephane Lavertu. 2018. "School Improvement Grants in Ohio: Effects on Student Achievement and School Administration." *Educational Evaluation and Policy Analysis* 40 (3):287-315.

Chiang, Hanley. 2009. "How Accountability Pressure on Failing Schools Affects Student Achievement." *Journal of Public Economics* 93 (9-10):1045-1057.

Chubb, John E., and Terry M. Moe. 1990. *Politics, Markets and America's Schools*. Washington, DC.: Brookings Institution Press.

Clark, Damon. 2009. "The Performance and Competitive Effects of School Autonomy." *Journal of Political Economy* 117 (4):745-783.

Colorado Department of Education. 2010. "Annual Report Concerning Districts of Innovation." https://www.cde.state.co.us/sites/default/files/documents/choice/download/sb130/annualreport2010.pdf.

Colorado Department of Education. 2013. "Colorado Innovation School Act." https://www.cde.state.co.us/sites/default/files/documents/choice/download/2013innovationreport_3.1.13.pdf.

Colorado Department of Education. 2015. 2015 Innovation Report. Colorado Department of Education.

Colorado Department of Education. n.d. "SchoolView." accessed June 22, 2019. https://www.cde.state.co.us/schoolview.

Dee, Thomas S. 2012. "Title." NBER Working Papers.

Dee, Thomas S., and Brian A. Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30 (3):418-446.

Dobbie, Will, and Roland G. Fryer. 2011. "Are High Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics* 3 (3):158-187.

Dragoset, Lisa, Jaime Thomas, Mariesa Herrmann, John Deke, Susanne James-Burdumy, Cheryl Graczewski, Andrea Boyle, Rachel Upton, Courtney Tanenbaum, and Jessica Giffin. 2017. School Improvement Grants: Implementation and Effectiveness. U.S. Department of Education.

Figlio, David, and Susanna Loeb. 2011. "School Accountability." In *Handbook of the Economics of Education*, edited by Eric A. Hanushek, Stephen Machin and Ludger Woessman, 383-421. The Netherlands: North-Holland: Elsevier.

Fryer, Roland G. 2014. "Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments." *The Quarterly Journal of Economics* 129 (3):1355-1407.

Hanushek, Eric. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19 (2):141-164.

Hanushek, Eric, and Margaret E. Raymond. 2005. "Does School Accountability Lead to Improved Student Performance." *Journal of Policy Analysis and Management* 24 (2):297-327.

Harris, Douglas N., and Matthew Larsen. 2016. The Effects of New Orleans Post-Katrina School Reforms on Student Academic Outcomes. Education Research Alliance for New Orleans.

Heissel, Jennifer A., and Helen F. Ladd. 2018. "School Turnaround in North Carolina: A Regression Discontinuity Analysis." *Economics of Education Review* 62 (302-320).

Imberman, Scott A. 2011. "The Effect of Charter Schools on Achievement and Behavior of Public School Students." *Journal of Public Economics* 95 (7-8):850-863.

Kraft, Matthew A. 2020. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher* 49 (4):241-253.

LaFortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. "School Finance Reform and the Distribution of Student Achievement." *American Economic Journal: Applied Economics* 10 (2):1-26.

Mackinnon, James G. 2016. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32 (2).

Mackinnon, James G., and Matthew D. Webb. 2018. "The Wild Bootstrap for Few (Treated) Clusters." *Econometrics Journal* 21:114-135.

Marsh, Julie A., Katherine o. Strunk, and Susan Bush. 2013. "Portfolio District Reform Meets School Turnaround: Early Implementation Findings from the Los Angeles Public School Choice Initiative." *Journal of Educational Administration* 51 (4):498-527.

Mora, Ricardo, and Iliana Reggio. 2017. "alternative Diff-in-Diffs estimators with Several Pre-Treatment Periods." *Econometrics Reviews* (forthcoming).

Robles, Yesenia. 2011. "Anxious Eyes on Denver's Far-Northeast Region as School Starts." *The Denver Post*, August 10, 2011. https://www.denverpost.com/2011/08/10/anxious-eyes-on-denvers-far-northeast-region-as-school-starts/.

Rockoff, Jonah, and Lesley J. Turner. 2010. "Short-Run Impacts of Accountability on School Quality." *American Economic Journal: Economics Policy* 2 (4):119-47.

Schueler, Beth E., Joshua S. Goodman, and David J. Deming. 2017. "Can States Take Over and Turn Around School Districts? Evidence from Lawrence, Massachusetts." *Educational Evaluation and Policy Analysis* 39 (2):311-332.

Steinberg, Matthew P. 2014. "Does Greater Autonomy Improve School Performance? Evidence from a Regression Discontinuity Analysis in Chicago." *Education Finance and Policy* 9 (1):1-35.

Strunk, Katherine O., Julie A. Marsh, Ayesha K. Hashim, Susan Bush-Mecenas, and Tracey Weinstein. 2016. "The Impact of Turnaround Reform on Student Outcomes: Evidence and Insights from the Los Angeles Unified School District." *Education Finance and Policy* 11 (3):251-282.

Sun, Min, Emily K. Penner, and Susanna Loeb. 2017. "Resource- and Approach-Driven Multidimensional Change: Three-Year Effects of School Improvement Grants." *American Educational Research Journal* 54 (4):607-643.

Tomassini, Jason. 2012. "Denver Turnaround Initiative Showing Achievement Gains." *Education Week*, June 5, 2012. https://www.edweek.org/ew/articles/2012/06/06/33blueprint.h31.html.

Yinger, John. 2004. *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity*. Cambridge, MA: MIT University Press.

Zimmer, Ron, Gary T. Henry, and Adam Kho. 2017. "The Effects of School Turnaround in Tennessee's Achievement School District and Innovation Zones." *Educational Evaluation and Policy Analysis* 39 (4):670-696.

Tables and Figures

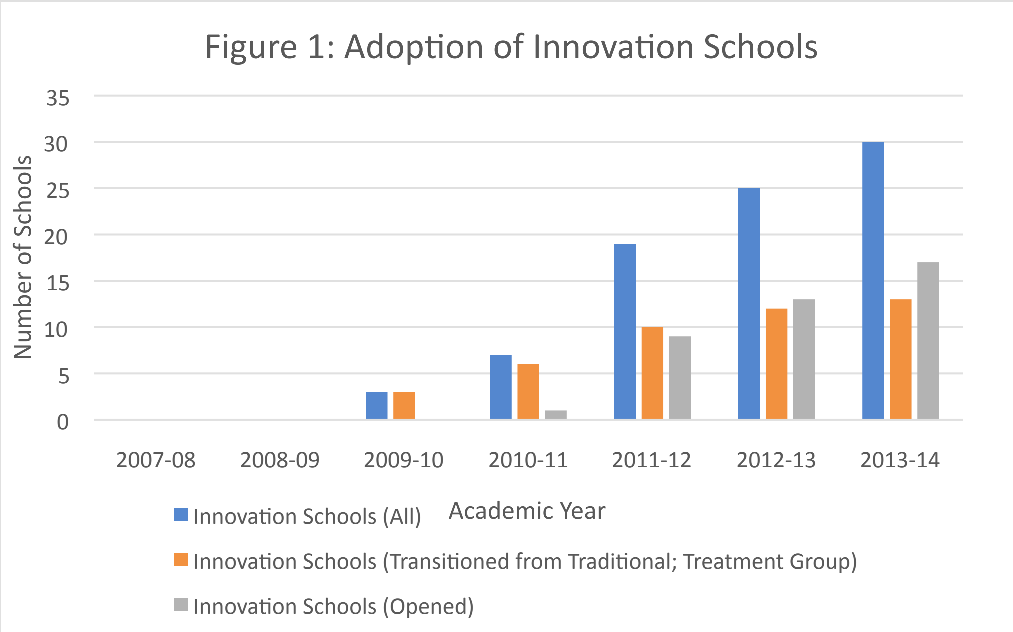


Table 1: Waivers Requested by DPS Innovation Schools (Colorado Department of Education 2013)	
Statutory Provision Waived	Percentage of Schools
Money	
Budgetary Control	100
Curriculum	
Educational Program and Textbook	76
Content Standards	80
Personnel	
Employee Performance Evaluations	92
Personnel Selection and Pay	100
Employee Dress Code	80
Principal Training	92
Termination of Personnel	96
Teacher Licensing	88
Teacher Contracts	92
Teacher Probation	92
Teacher Transfers	92
Grounds for Dismissal	96
Procedure for Dismissal	88
Teacher Salaries	92

Time	
Longer School Day and School Year	100

Table 2: Descriptives

Variable	Obs	Mean	Std. Dev.	Min	Max
Writing	1,015	491.57	40.33	419.88	624.56
Writing Z	1,015	0.00	1.00	-1.72	3.21
Reading	1,015	583.85	47.01	493.58	704.82
Reading Z	1,015	0.00	1.00	-2.08	2.57
Math	1,015	490.11	52.45	390.50	642.37
Math Z	1,015	0.00	1.00	-1.95	2.91
Innovation Schools	1,015	0.04	0.20	0.00	1.00
Enrollment	1,015	532.88	314.12	42.00	2435.00
Student Teacher Ratio	1,015	17.45	8.57	3.37	264.00
Teacher Salary	901	52947.73	7471.89	31093.52	82848.02
Discipline	904	47.94	68.47	0.00	494.00
% Free Lunch	1,015	0.71	0.27	0.03	1.00
% Black	1,015	0.14	0.14	0.00	0.86
% Hisp	1,015	0.58	0.28	0.04	0.98

Table 3: Comparison Between Innovation and Non-Innovation Schools in 3-year Pre-Treatment Period (2006-07 to 2008-09)

	Non-Innovation Schools (NIS) (105 Schools)	PSM Matches (PSM) (44 Schools)	Innovation Schools (IS) (12 Schools)	p-value (NIS vs. IS)	p-value (PSM vs. IS)
N	308	130	31		
Writing	486.67 (40.53)	462.28 (28.74)	466.11 (30.65)	0.006	0.51
Writing Z	0.05 (1.01)	-0.56 (0.71)	-0.47 (0.76)	0.006	0.54
Reading	579.63 (49.03)	549.89 (38.38)	550.92 (37.30)	0.002	0.89
Reading Z	0.05 (1.00)	-0.55 (0.78)	-0.54 (0.76)	0.002	0.91
Math	483.80 (50.32)	451.97 (37.73)	453.20 (40.71)	0.001	0.87
Math Z	0.06 (1.00)	-0.58 (0.75)	-0.55 (0.80)	0.001	0.89
Enrollment	534.10 (320.83)	533.59 (329.19)	514.00 (238.51)	0.73	0.76
Student Teacher	18.46 (14.50)	17.28 (3.99)	16.80 (2.53)	0.52	0.53

Ratio					
Teacher Salary (Thousands)	54.77 (69.39)	54.72 (51.60)	53.11 (46.88)	0.28	0.19
Discipline	55.93 (84.35)	59.17 (86.17)	77.18 (100.84)	0.27	0.40
% Free Lunch	0.65 (0.27)	0.81 (0.16)	0.78 (0.09)	0.006	0.44
% Black	0.15 (0.15)	0.15 (0.18)	0.21 (0.19)	0.030	0.082
% Hispanic	0.55 (0.30)	0.69 (0.23)	0.63 (0.21)	0.14	0.21

Table 4: Effects of Innovation Schools on Academic Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Writing	Writing	Reading	Reading	Math	Math
Innovation Schools	0.111+	0.108	0.136*	0.113+	0.232*	0.225*
	(0.0669)	(0.0660)	(0.0649)	(0.0598)	(0.0969)	(0.0970)
	0.146	0.0135	0.062+	0.095+	0.043*	0.041*
Full Sample	x		x		x	
Matched Sample		x		x		x
School FE	x	x	x	x	x	x
Year FE	x	x	x	x	x	x
Observations	1,015	464	1,015	464	1,015	464
R-squared	0.083	0.090	0.086	0.134	0.092	0.084
Number of Panels	148	62	148	62	148	62

Cluster robust standard errors in parentheses
Wild bootstrap p-values in third row
** p<0.01, * p<0.05, + p<0.1

Table 5: Event-study Regressions (Year Negative One Omitted)

VARIABLES	(1) Writing	(2) Writing	(3) Reading	(4) Reading	(5) Math	(6) Math
Year Negative Seven	0.0362	0.0531	0.0979	0.188*	-0.0171	-0.0250
	(0.0583)	(0.0724)	(0.0758)	(0.0842)	(0.0842)	(0.0996)
Year Negative Six	0.190*	0.214*	0.0817	0.152	-0.0878	-0.0760
	(0.0908)	(0.0989)	(0.140)	(0.132)	(0.142)	(0.143)
Year Negative Five	0.0318	0.0269	-0.0329	-0.0185	-0.191	-0.163
	(0.0883)	(0.0873)	(0.136)	(0.136)	(0.118)	(0.123)
Year Negative Four	0.0261	0.0254	-0.0141	0.00930	0.0258	0.0370
	(0.0673)	(0.0715)	(0.0857)	(0.0815)	(0.0862)	(0.0899)
Year Negative Three	-0.0648	-0.0697	-0.0121	0.00958	-0.0155	-0.0127
	(0.0794)	(0.0836)	(0.0953)	(0.0980)	(0.106)	(0.113)
Year Negative Two	-0.00993	-0.0169	-0.128+	-0.121	-0.0117	0.000351
	(0.0441)	(0.0450)	(0.0759)	(0.0736)	(0.0903)	(0.0935)
Year Negative One	(Omitted)	(Omitted)	(Omitted)	(Omitted)	(Omitted)	(Omitted)
Year One	0.0311	0.0256	0.0333	0.0271	0.129	0.135
	(0.0501)	(0.0528)	(0.0551)	(0.0563)	(0.0961)	(0.0971)
Year Two	0.218**	0.215**	0.172*	0.164*	0.310*	0.304*
	(0.0695)	(0.0703)	(0.0767)	(0.0741)	(0.120)	(0.123)
Year Three	0.128	0.127	0.111	0.0995	0.284+	0.273+
	(0.125)	(0.126)	(0.141)	(0.138)	(0.155)	(0.159)
Year Four	0.0552	0.0342	0.128	0.107	0.175	0.189
	(0.138)	(0.140)	(0.155)	(0.159)	(0.187)	(0.195)
Year Five	-0.0537	-0.0718	0.0421	0.0122	-0.0135	-0.00280
	(0.143)	(0.147)	(0.175)	(0.195)	(0.211)	(0.221)
Full Sample	x		x		x	
Matched Sample		x		x		x
School FE	x	x	x	x	x	x
Year FE	x	x	x	x	x	x
Observations	1,015	464	1,015	464	1,015	464
R-squared	0.094	0.119	0.094	0.154	0.103	0.106
Number of Panels	148	62	148	62	148	62

Cluster robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

Table 6: Heterogenous Effects of Innovation Schools on Academic Outcomes (By Effect Size and Direction)

	Full Sample	Matched Sample	Full Sample	Matched Sample
Writing	0.228**	0.224**	-0.136	-0.152+
	(0.0653)	(0.0658)	(0.0836)	(0.0801)
Reading	0.254**	0.222**	-0.116	-0.132
	(0.0614)	(0.0551)	(0.0991)	(0.101)
Math	0.419**	0.408**	-0.160+	-0.168+
	(0.0919)	(0.0894)	(0.0947)	(0.0927)
High Performers (8 Schools)	x	x		
Low Performers (5 Schools)			x	x
School FE	x	x	x	x
Year FE	x	x	x	x
Observations	980	429	955	404
Number of Panels	143	57	140	54

Cluster robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

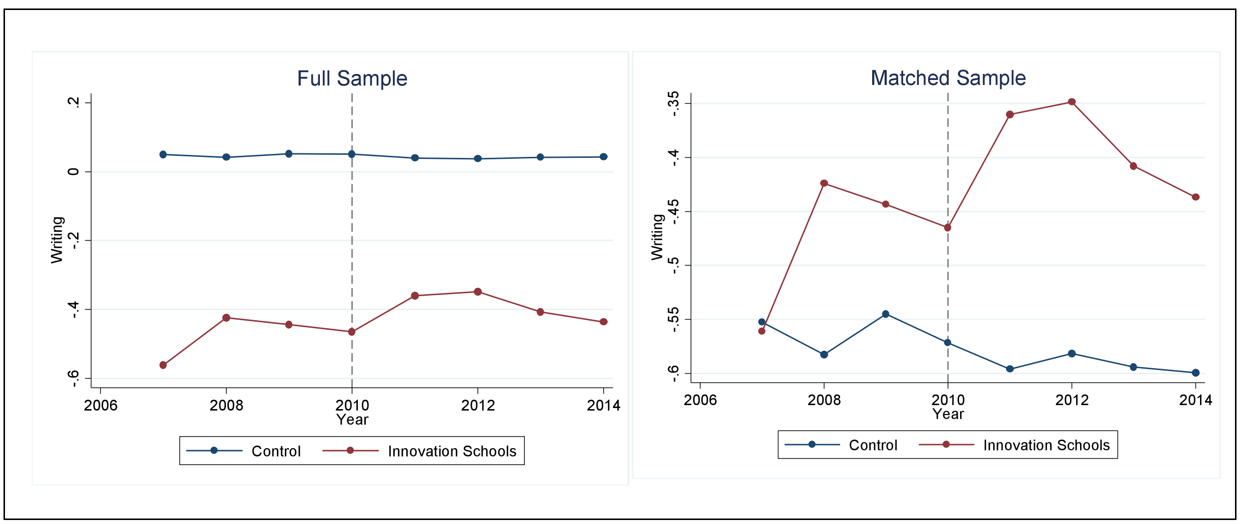
Table 7: Heterogenous Effects of Innovation Schools on Academic Outcomes (By External Management and SIG Funding)

	Full Sample	Matched Sample	Full Sample	Matched Sample
Writing	0.197*	0.194*	0.216*	0.211*
	(0.0827)	(0.0816)	(0.0939)	(0.0940)
Reading	0.255**	0.205**	0.297**	0.239**
	(0.0818)	(0.0712)	(0.0845)	(0.0754)
Math	0.302**	0.297**	0.322**	0.321**
	(0.0683)	(0.0749)	(0.0757)	(0.0833)
Without DSSN	x	x		
Without DSSN & SIG			x	x
School FE	x	x	x	x
Year FE	x	x	x	x
Observations	964	413	958	407
Number of Panels	141	55	140	54

Cluster robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

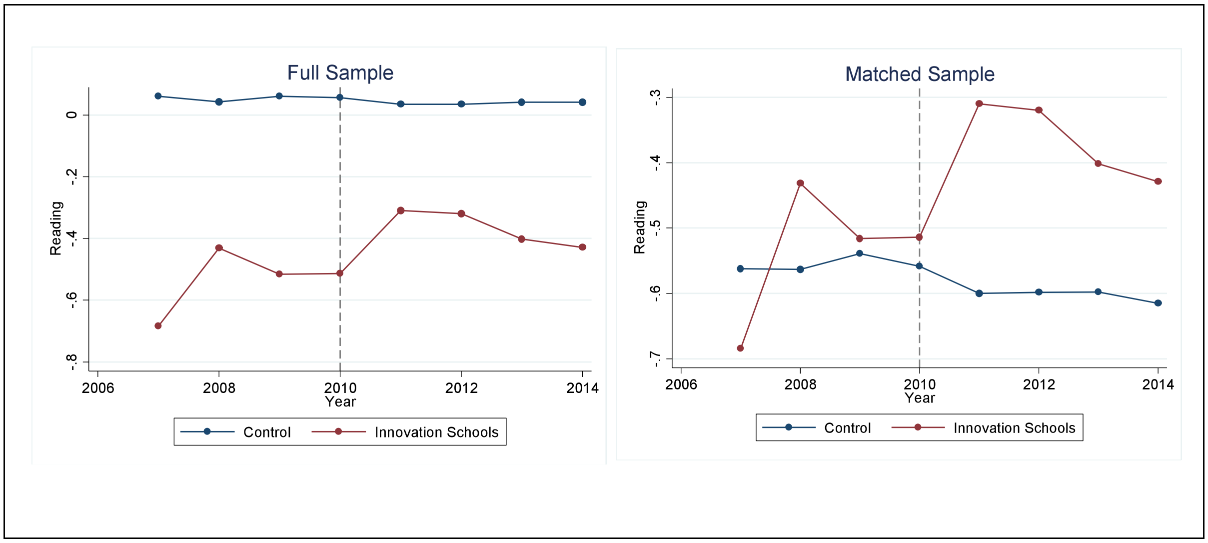
Table 8: Effects of Innovation Schools on Organizational Outcomes		
	Full Sample	Matched Sample
Enrollment	17.58	9.894
	(16.76)	(18.06)
Student-Teacher Ratio	1.632	0.371
	(1.339)	(0.515)
Teacher Salary	-2,475*	-2,477*
	(1,117)	(1,057)
Discipline	-8.961	-8.606
	(10.94)	(11.42)
All Innovation Schools	x	x
School FE	x	x
Year FE	x	x
Observations	904	412
Number of Panels	148	62
Cluster robust standard errors in parentheses		
** p<0.01, * p<0.05, + p<0.1		

Figure 2: Trend Graphs (Writing Scores, Main Sample)



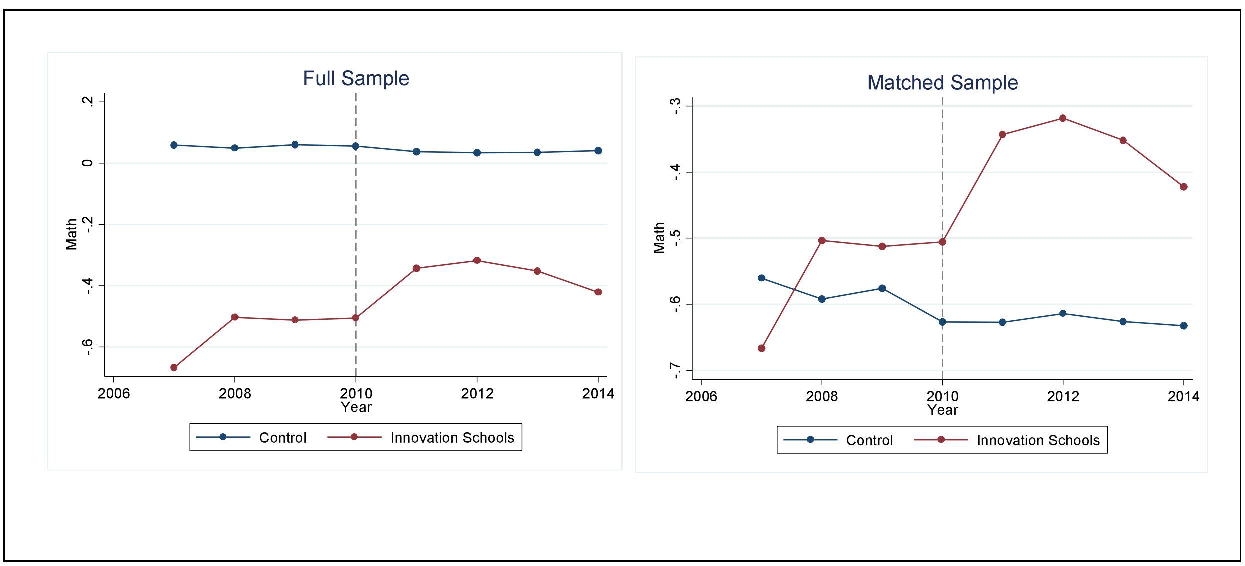
Note. These figures are binned scatter plots. This graph plots average Writing Score for treatment and control schools from 2007 and 2014.

Figure 3: Trend Graphs (Reading Scores, Main Sample)



Note. These figures are binned scatter plots. This graph plots average Reading score for treatment and control schools from 2007 and 2014.

Figure 4: Trend Graphs (Math Scores, Main Sample)



Note. These figures are binned scatter plots. This graph plots average Math score for treatment and control schools from 2007 and 2014.

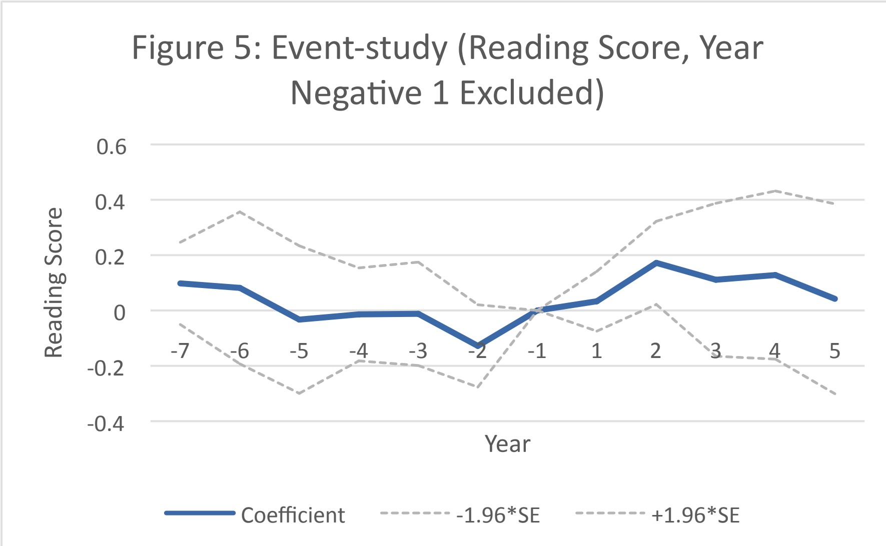


Figure 5: Event-study (Reading Score, Year Negative 1 Excluded)
Reading Score
0.6
0.4
0.2
0
-0.2
-0.4
-7 -6 -5 -4 -3 -2 -1 1 2 3 4 5
Year
Coefficient
-1.96*SE
+1.96*SE

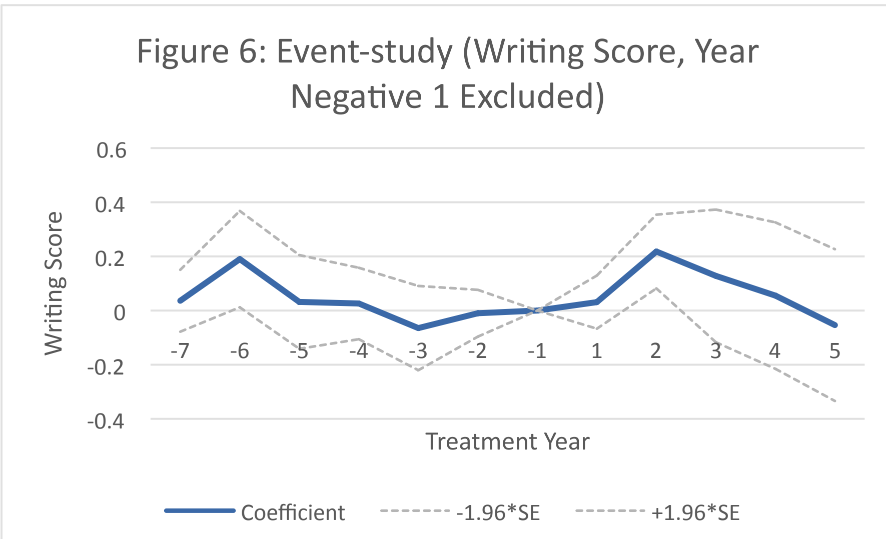


Figure 6: Event-study (Writing Score, Year Negative 1 Excluded)
Writing Score
0.6
0.4
0.2
0
-0.2
-0.4
-7 -6 -5 -4 -3 -2 -1 1 2 3 4 5
Treatment Year
Coefficient
-1.96*SE
+1.96*SE

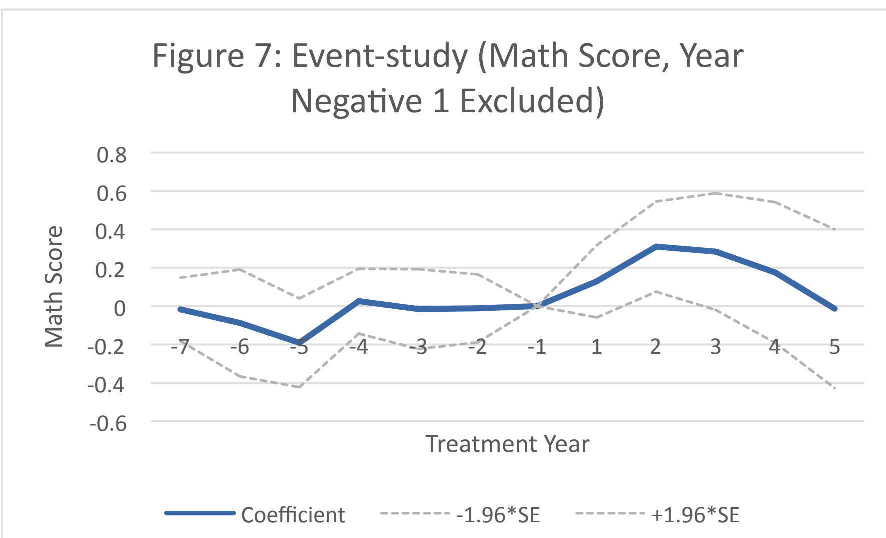


Figure 7: Event-study (Math Score, Year Negative 1 Excluded)
Math Score
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-7 -6 -5 -4 -3 -2 -1 1 2 3 4 5
Treatment Year
Coefficient
-1.96*SE
+1.96*SE

Appendix Tables:

Appendix Table 1: Student Sorting Models (Treatment Effects on Demographic Controls)

VARIABLES	(1) % Free Lunch	(2) % Free Lunch	(3) % Black	(4) % Black	(5) % Hisp	(6) % Hisp
Innovation Schools	0.00574	- 0.00575	-0.0118	-0.0154	-0.0126	-0.0180
	(0.0202)	(0.0217)	(0.0105)	(0.00980)	(0.0103)	(0.0132)
Full Sample	x		x		x	
Matched Sample		x		x		x
School FE	x	x	x	x	x	x
Year FE	x	x	x	x	x	x
Observations	1,015	464	1,015	464	1,015	464
R-squared	0.403	0.432	0.409	0.566	0.420	0.534
Number of Panels	148	62	148	62	148	62

Cluster robust standard errors in parentheses

** $p<0.01$, * $p<0.05$, + $p<0.1$

Appendix Table 2: Placebo Tests

VARIABLES	(1) Writing	(2) Writing	(3) Reading	(4) Reading	(5) Math	(6) Math
Placebo 2008	-0.0001	0.0248	0.109	0.0722	0.0417	0.0504
	(0.0502)	(0.0553)	(0.0682)	(0.0856)	(0.0617)	(0.0686)
Placebo 2009	0.0383	0.00840	0.0827+	0.0212	0.0924	0.103
	(0.0500)	(0.0599)	(0.0428)	(0.0588)	(0.0679)	(0.0738)
Placebo 2010	-0.00397	-0.0212	0.0540	-0.0173	0.0678	0.0825
	(0.0614)	(0.0719)	(0.0925)	(0.0986)	(0.0845)	(0.0993)
Placebo 2011	0.0117	0.0200	0.120	0.0723	0.162+	0.175+
	(0.0807)	(0.0863)	(0.0848)	(0.0944)	(0.0869)	(0.103)
Placebo 2012	-0.0906	-0.0842	0.00131	-0.0412	0.0987	0.118
	(0.0590)	(0.0679)	(0.0659)	(0.0752)	(0.0740)	(0.0814)
Placebo 2013	-0.285**	-0.291**	-0.361**	-0.432**	-0.240**	-0.264**
	(0.0490)	(0.0661)	(0.0525)	(0.0758)	(0.0608)	(0.0882)
Full Sample	x		x		x	
Matched Sample		x		x		x
School FE	x	x	x	x	x	x
Year FE	x	x	x	x	x	x
Observations	837	371	837	371	837	371
R-squared	0.075	0.087	0.076	0.116	0.089	0.085
Number of Panels	143	62	143	62	143	62

Cluster robust standard errors in parentheses

** p<0.01, * p<0.05, + p<0.1

Appendix Table 3: Event-study Regressions (Year Negative Three Omitted)

VARIABLES	(1) Writing	(2) Writing	(3) Reading	(4) Reading	(5) Math	(6) Math
Year Negative Seven	0.101+	0.123*	0.110*	0.179**	-0.00159	-0.0123
	(0.0522)	(0.0598)	(0.0470)	(0.0559)	(0.0597)	(0.0702)
Year Negative Six	0.255**	0.283**	0.0938	0.142	-0.0723	-0.0633
	(0.0760)	(0.0825)	(0.106)	(0.0980)	(0.108)	(0.110)
Year Negative Five	0.0966	0.0967+	-0.0208	-0.0280	-0.175**	-0.150**
	(0.0589)	(0.0565)	(0.0782)	(0.0816)	(0.0489)	(0.0529)
Year Negative Four	0.0910	0.0951	-0.00196	-0.000273	0.0413	0.0496
	(0.0602)	(0.0636)	(0.0550)	(0.0522)	(0.0590)	(0.0588)
Year Negative Three	(Omitted)	(Omitted)	(Omitted)	(Omitted)	(Omitted)	(Omitted)
Year Negative Two	0.0549	0.0528	-0.116+	-0.131+	0.00377	0.0130
	(0.0571)	(0.0583)	(0.0688)	(0.0769)	(0.0773)	(0.0778)
Year Negative One	0.0648	0.0697	0.0121	-0.00958	0.0155	0.0127
	(0.0794)	(0.0836)	(0.0953)	(0.0980)	(0.106)	(0.113)
Year One	0.0959	0.0954	0.0454	0.0176	0.144	0.148
	(0.0888)	(0.0965)	(0.0906)	(0.0945)	(0.128)	(0.133)
Year Two	0.283**	0.285**	0.184*	0.154*	0.326*	0.316*
	(0.0875)	(0.0920)	(0.0829)	(0.0750)	(0.130)	(0.133)
Year Three	0.193+	0.197+	0.123	0.0899	0.299*	0.286*
	(0.115)	(0.117)	(0.131)	(0.127)	(0.120)	(0.128)
Year Four	0.120	0.104	0.140	0.0970	0.191	0.201
	(0.124)	(0.124)	(0.142)	(0.134)	(0.167)	(0.177)
Year Five	0.0112	-0.00206	0.0542	0.00266	0.00201	0.00986
	(0.137)	(0.137)	(0.170)	(0.172)	(0.191)	(0.201)
Full Sample	x		x		x	
Matched Sample		x		x		x
School FE	x	x	x	x	x	x
Year FE	x	x	x	x	x	x
Observations	1,015	464	1,015	464	1,015	464
R-squared	0.094	0.119	0.094	0.154	0.103	0.106
Number of Panels	148	62	148	62	148	62

Cluster robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

Appendix Table 4: School-Specific Linear Time Trends (Main Models)						
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Writing	Writing	Reading	Reading	Math	Math
Innovation School	0.135+	0.122+	0.124	0.119	0.206+	0.194+
	(0.0739)	(0.0718)	(0.0846)	(0.0768)	(0.117)	(0.115)
Full Sample	x		x		x	
Matched Sample		x		x		x
School FE	x	x	x	x	x	x
Year FE	x	x	x	x	x	x
School-Specific Trends	x	x	x	x	x	x
Observations	1,015	464	1,015	464	1,015	464
R-squared	0.469	0.440	0.482	0.470	0.531	0.495
Number of Panels	148	62	148	62	148	62
Cluster robust standard errors in parentheses						
** p<0.01, * p<0.05, + p<0.1						

Appendix Table 5: Lagged Dependent Variable Models (Main Models)						
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Writing	Writing	Reading	Reading	Math	Math
Innovation School	0.0884	0.0867	0.0711	0.0767	0.177*	0.179*
	(0.0538)	(0.0550)	(0.0498)	(0.0504)	(0.0694)	(0.0728)
Full Sample	x		x		x	
Matched Sample		x		x		x
School FE	x	x	x	x	x	x
Year FE	x	x	x	x	x	x
School-Specific Trends	x	x	x	x	x	x
Observations	865	402	865	402	865	402
R-squared	0.205	0.186	0.219	0.193	0.290	0.237
Number of Panels	142	62	142	62	142	62
Cluster robust standard errors in parentheses						
** p<0.01, * p<0.05, + p<0.1						

Figure 8: Event-study (Writing Score, Year Negative 3 Excluded)

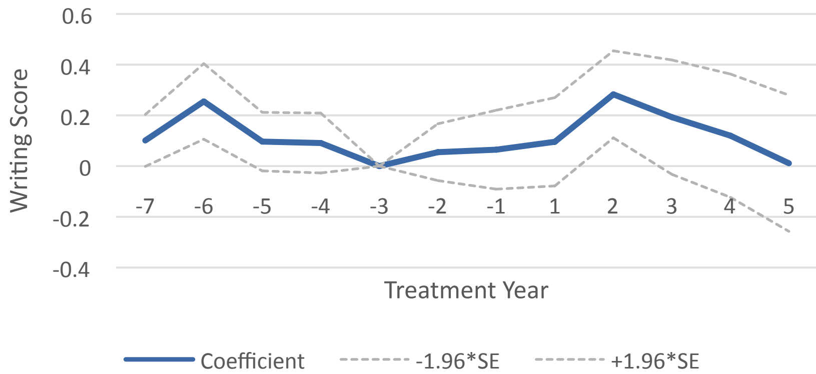


Figure 9: Event-study (Reading Score, Year Negative 3 Excluded)

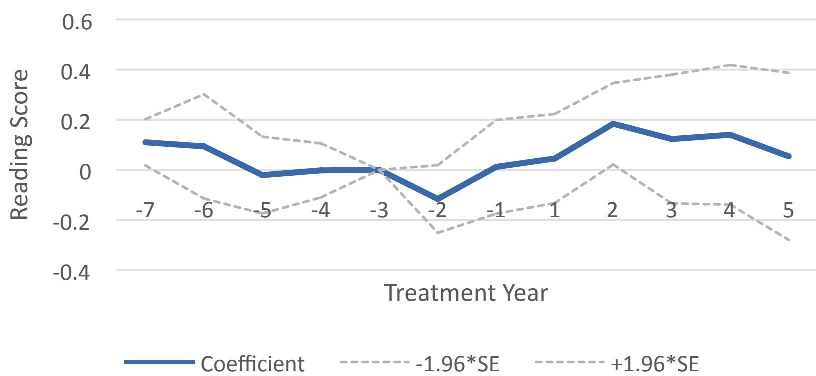


Figure 10: Event-study (Math Score, Year Negative 3 Excluded)