-- PRELIMINARY; DO NOT QUOTE; COMMENTS WELCOMED --

The Effects of the New Orleans Post-Katrina School Reforms on Student Academic Outcomes

Douglas N. Harris
Matthew F. Larsen

November 2, 2015

Abstract: The school reforms put in place in New Orleans after Hurricane Katrina represent the most intensive test-based and market-based school accountability system ever created in the United States. Collective bargaining was ended, yielding flexible human capital management. Traditional attendance zones were eliminated, expanding choice for families. And almost all public schools were taken over by the state, which turned over management to outside non-profit charter management organizations working under performance contracts. Ten years later, this study provides the first examination of the effects of this package of reforms on student achievement. Identification is based on multiple difference-in-difference (DD) strategies, using outcomes before and after the hurricane and reforms in New Orleans and a matched comparison group that experienced hurricane damage but not the school reforms. The estimation procedures address potential threats to identification, including changes in the population, strategic behavior in test scores from high-stakes accountability, the influence of the interim schools attended by evacuated students, and the trauma and disruption from the hurricane itself. With the possible exception of test-based accountability strategic behavior, these factors seem to have a small influence and, collectively, they appear to cancel each other out. The results suggest that, over time, as the reforms yielded a new system of schools, they had large positive cumulative effects of 0.2-0.4 standard deviations.

Acknowledgements: This study was conducted at the Education Research Alliance for New Orleans at Tulane University. The authors wish to thank the organization's funders: the John and Laura Arnold Foundation, William T. Grant Foundation, the Spencer Foundation and, at Tulane, the Department of Economics, Murphy Institute and School of Liberal Arts. We also thank Joshua Angrist, Robert Bifulco, Joshua Cowen, John Easton, Adam Gamoran, Dan Goldhaber, Helen Ladd, Susanna Loeb, Parag Pathak, Ross Rubenstein, Robert Santillano, Amy Schwartz, John Yinger, Lindsey Bell Weixler and seminar participants at Syracuse University, Tulane University, the University of Arkansas, and the University of Virginia. For other important contributions we thank Alica Gerry and Nathan Barrett.

Author Information: Douglas N. Harris (corresponding author) is Professor of Economics, Schleider Foundation Chair in Public Education, and Director of the Education Research Alliance for New Orleans at Tulane University (dharri5@tulane.edu). Matthew F. Larsen is an Assistant Professor of Economics at Lafayette College and a Research Associate at ERA-New Orleans (larsenmf@lafayette.edu).

Introduction

For the past century, America's publicly funded schools have been almost universally operated by local government agencies that assign students to schools based on their neighborhoods. This type of system could generate competition among school districts and yield an efficient equilibrium (Tiebout, 1956), though this might not occur in the presence of political forces (Kollman, Miller, & Page, 1997), labor unions (Hoxby, 1996; Strunk & Grissom, 2010), and other factors that may make public sector production inefficient (Hill, Pierce, & Guthrie, 1997; Chubb & Moe, 1990).[1] For these and other reasons, Friedman (1962) argued that families should be "free to choose" where their children attend school, government subsidies should follow the student to induce more direct competition among schools, and non-governmental suppliers should be allowed into the market through performance-based contracts that give them autonomy over how objectives are reached.

The school reforms put in place in New Orleans after the tragedy of Hurricane Katrina offer arguably the first direct test of these two alternative models. Prior to Katrina, the New Orleans school system was well aligned with almost every other city in the United States. In addition to neighborhood-based assignment of students to schools, the vast majority of schools were operated by the local school district, the New Orleans Public Schools, and governed by a locally elected body, the Orleans Parish School Board (OPSB). Teachers worked under union contracts that established single salary schedules and work rules.

[1] The Tiebout analysis is usually based on between-district competition, which did not change in New Orleans. The larger issue, however, is whether Tiebout-type competition generates efficient equilibria or whether additional market mechanisms might improve efficiency. See Bewell (1981) for a more skeptical theoretical examination of Tiebout.

[2] Hurricane Rita struck just one month later on September 24, 2005. For simplicity, however, we simply refer to "the hurricane" going forward.

After Hurricane Katrina struck New Orleans on August 29, 2005, all the hallmarks of the traditional school district had been eliminated.[2] The state government took over the school system, moving oversight of almost all the city's public schools from the local OPSB to the statewide Louisiana Recovery School District (RSD). Many OPSB schools were quickly turned into charter schools and, over time, so too were all RSD schools. Attendance zones were eliminated, creating open school choice for families. All educators were fired. The teacher union contract was allowed to expire and never replaced. Local and state agencies still had a role, especially in funding schools, but they no longer exercised much control, except in passing funds on to schools on a per-pupil basis and deciding which schools would be opened and closed. In short, over just a few years, the government role was dramatically altered and reduced, from operator to oversight body. The "one best system" of U.S. public education (Tyack, 1974) was eliminated for the first time in a century.

As sudden as these changes were in New Orleans, the new policies themselves reflected a two-decade shift toward test-based and market-based accountability throughout the United States. Induced by evidence of possibly inefficient resource use (Hanushek, 1996), poor showings on international assessments (National Commission on Educational Excellence, 1983; Goldin & Katz, 2008) and flat test score trends (Hanushek & Woessman, 2010), the federal Elementary and Secondary Schools Act (ESEA) began requiring standardized testing and school report cards (Harris & Herrington, 2006). Under the most recent incarnation of ESEA, known as *No Child Left Behind* (NCLB), the government also increased the frequency and stakes attached to those test scores (Dee &

[2] Hurricane Rita struck just one month later on September 24, 2005. For simplicity, however, we simply refer to "the hurricane" going forward.

Jacob, 2011). The source of accountability was still within the government, but with incentives akin to performance-based contracting. The New Orleans reforms also followed the longer national trend toward market accountability through parental school choice and opening up the supply side through charter schools (Angrist, Pathak & Walters, 2011), private school vouchers (Rouse, 1998; Krueger & Zhu, 2004), and intra- and inter-district choice among traditional public schools (Harris & Witte, 2011). With accountability from both the government contracts and markets, the theory is that leaders would have incentives to perform, and autonomy to meet accountability demands, yielding greater efficiency.

Though the word accountability has been commonly used, the actual incentives have been weaker than advocates desired. Some districts around the country had experimented with school-level autonomy (Ravitch, 2000) and mayoral and state takeovers (Wong & Shen, 2006; Gill et al., 2006), but most of these efforts were short-lived and the influence of local school board politics, labor unions, and school attendance zones still dominated school operations (Ravitch, 2000). NCLB increased the volume of testing and changed school practices (Rouse et al., 2013), but only a small fraction of the schools slated for corrective action under NCLB experienced significant intervention (GAO, 2007).[3] This may be why researchers have found the NCLB effects to be so small (Dee & Jacob, 2011).

The same could be said of market accountability. Only two percent of U.S. students attend charter schools and 13 percent of U.S. students attend a non-assigned

[3] A synthesis of evidence from studies of pre-NCLB accountability found more positive cumulative effects on test scores averaging 0.08 standard deviations (Lee, 2008). Also, see Carnoy & Loeb (2003).

publicly funded school (Harris & Witte, 2011).[4] At the time of Hurricane Katrina, only seven districts had more than 20 percent of their students in charter schools and none were above 50 percent (National Alliance for Public Charter Schools, 2013). While research is increasingly showing positive effects of charter schools (e.g., Angrist et al., 2010, 2011a, 2011b, forthcoming; CREDO, 2013a, 2013b), their market share has been too small to affect outcomes across entire cities or regions, or to generate competitive effects on nearby on traditional public schools (Gill & Booker, 2008; Epple, Romano & Zimmer, 2015).[5] For these reasons, advocates for accountability and school autonomy have argued that policymakers have not gone far enough (Hill & Lake, 2004, Evers, 2014; Peterson, 2014; Walberg, 2014).

In New Orleans, policymakers went as far with accountability as one could imagine. However, the evidence on this remarkable post-Katrina policy experiment has been quite limited. Most of the debate centers on positive upward trends in outcomes (Cowen Institute, 2013). New Orleans statewide ranking on the percentage of students who are proficient has moved from the 67th ranked district to the 39th (of 68) ranked districts since the hurricanes (Louisiana Department of Education, 2015).[6] Figures 1A-1H reinforce the idea that significant improvement occurred. Averaging across all subjects and grades, we find that the test score gap between New Orleans and the rest of the state decreased by 0.34 standard deviations from 2004-05 to 2011-12 (see Table 1). These positive trends, combined with evidence that charter schools in New Orleans

[4] We include in "non-assigned publicly funded schools" students who attend schools labeled charter, magnet, and intra-and inter-district schools of choice (Harris & Witte, 2011).
[5] Among all the studies that have examined the competitive effects of charter schools and vouchers on traditional public schools, about half find evidence of such effects on student test scores (Gill & Booker, 2008). Other studies have examined the effects of competition within the traditional public schooling market and these too are mixed (e.g., Hoxby, 2000; Belfield & Levin, 2003; Rothstein, 2007).
[6] For comparability, the post-Katrina New Orleans "district" ranking is based on a weighted average of the RSD and OPSB schools.

(Abdulkadiroğlu et al., 2015) and Louisiana (CREDO, 2013a, 2013b) are more effective than traditional public schools, suggest the reform effects probably have been positive.

With these positive signs, the system has been widely hailed among reform advocates (e.g., Whitehurst, 2012) and national political leaders with otherwise divergent views, from Democratic President Obama (2010) and his Education Secretary Arne Duncan to Republican Louisiana Governor and presidential candidate Bobby Jindal (America Next, 2015). Also, at least 27 districts are following New Orleans's lead (Hill & Campbell, 2011).

Unfortunately, the evidence to date provides little evidence of the effectiveness of the New Orleans reform package. The studies to date (Abdulkadiroğlu et al., 2015; CREDO, 2013a, 2013b) have focused entirely on the post-Katrina period and are therefore not focused on the effects of the post-Katrina change in policy.[7] In this study, we use several difference-in-difference strategies comparing the pre- and post-reform periods in New Orleans relative to matched comparison groups. The results suggest that the school reforms had a cumulative achievement effect of 0.2-0.4 standard deviations (8-15 percentile points[8]) seven years after the reforms. While the effect magnitudes were much smaller than this at first, they grew steadily and the long-term effects are generally statistically significant. We can also largely rule several threats to identification, including population change, trauma and disruption from the hurricane, the effectiveness of the interim schools that evacuated students attended temporarily, and strategic

[7] The Center for Research on Education Outcomes (CREDO, 2013a, 2013b) compared student growth in New Orleans with growth of similar students ("virtual twins") in traditional public schools in other districts, all in the post-Katrina period. Also, Sacerdote (2012) finds that New Orleans evacuees experienced larger increases in school quality than evacuees from other Louisiana parish/districts, which confirms the low performance of pre-Katrina New Orleans schools, but he does not address their post-Katrina improvement.

[8] This is based off of students starting at the 50th percentile, which is the typical translation, though note that New Orleans students were at about the 30th percentile after Katrina.

behavior from test-based accountability. The treatment effects appear to be an order of magnitude larger than the potential biases, and some biases appear to cancel out.

This study highlights the long-term potential of intensive market- and test-based school reform. The next section describes threats to identification and our empirical strategies for addressing them. This is followed by discussion of data, results, and conclusions.

Model and Identification

Threats to Identification

There are many general threats to identification with natural experiments, including that policy adoption is endogenous. In the case of the New Orleans school reforms, we have five key additional threats; that is, there many alternative potential causes of the changes in measureable outcomes shown in Figures 1A-1H.

First, the population of the city changed (The Data Center, 2014; Vigdor, 2008). City leaders decided in the process of rebuilding the city to shut down and eventually replace most of major public housing projects. For this and other reasons, low-income residents may not have returned and this by itself could have increased scores in the city.

Second, when Louisiana families evacuated, they generally placed their children, temporarily, in the public schools in the cities to which they evacuated. There is evidence that New Orleans evacuees experienced larger gains in school quality in these “interim schools” relative to non-New Orleans evacuees (Sacerdote, 2012). If these gains did not fade out[9], then some of the later increases in achievement observed when students

[9] See, for example, McCaffrey et al. (2009) who study the fade out of teacher value-added over time.

returned to New Orleans might reflect the performance of these interim schools rather than the New Orleans reforms.

Third, prior research has shown that schools manipulate high-stakes measures and/or reallocate resources in ways that reduce lower-stakes measures (Figlio, 2006; Jacob, 2005; Koretz, 2009). Such strategic behavior may be especially important in New Orleans where schools are closed based substantially on test scores (Ruble & Harris, 2015) and accountability pressures are generally high.

Fourth, NCLB had been adopted a few years prior to Katrina and the law's key provisions were about to be implemented. Since the federal law focuses on low-performing schools, and since low-performing schools are the focus of NCLB sanctions, the post-Katrina improvements in outcomes might have occurred anyway.

While these first four threats to identification suggest the trends would tend to bias estimated effects upwards, the direction of the fourth threat could have the opposite influence: Hurricane Katrina was one of the worst disasters in American history[10] and created trauma (DeSalvo et al., 2007) and anxiety (Elliott & Pais, 2006) for residents that persisted many years later (Weems et al., 2010). Some of these psychological effects were driven by poor labor market outcomes among those who had lived in the most heavily flooded areas (Groen & Polivka, 2008). Those with worse post-hurricane housing and labor market outcomes also experienced worse Post-Traumatic Stress Disorder (PTSD) (Elliott & Pais, 2006). While most of the psychological evidence pertains to adults, there is also evidence of trauma and disruption among children more than two

[10] As many as 1,900 people died as a result of the storm and the city experienced at least $80 billion dollars in damage to physical infrastructure (Pane et al., 2008).

years after the hurricanes (Brown et al., 2011)[11] and this apparently reduced academic learning at least in the short term (Pane et al., 2006, 2008; Sacerdote, 2012).

Estimation Strategy

We use difference-in-difference (DD) analysis to address all these threats. Specifically, we estimate the effects of the New Orleans school reform package starting with standard two-period difference-in-difference estimation (Angrist & Pischke, 2009):

$$A_{jt} = \gamma_j + \lambda d_t + \delta(NOLA \cdot d_t) + \varepsilon_{jt} \qquad (1)$$

where A_{jt} is the achievement of students in school district j at time t, γ_j is a vector of group (school district) fixed effects, d_t indicates whether the outcomes pertain to a single pre-treatment period or a single post-treatment period, and $NOLA$ is an indicator set to unity for New Orleans and zero for all districts in the comparison group. Under certain assumptions, especially that student outcomes would have moved in parallel absent the treatment, ordinary least squares estimation of δ provides an unbiased estimate of the average treatment effect.[12]

Our first estimates are based on the estimation of equation (1) using only the year prior to the reforms (2004-05) and the most recent post-reform period available in the data (2008-09 or 2011-12, depending on the analysis). An alternative would be to average the post-reform periods together. However, this is inappropriate in this case because there are reasons to expect dynamic effects. In creating an entirely new system of schooling, New Orleans,leaders not only had to create new schools, but an entirely new governance

[11] One sample of students reported thoughts of the following common disaster-related events 30 months after the hurricanes: "having thoughts someone might die (79%), having clothes or toys ruined (78%), having their home badly damaged or destroyed (65%), witnessing others hurt during the storm (45%), having a pet hurt or die (41%), thinking they might die during the storm (38%), having trouble getting food and water (20%)." (Brown et al. 2011, p.576).

[12] Athey and Imbens (2002) discuss the linearity assumptions used in DD estimation.

structure and new institutions to recruit and develop charter school operators (New Schools for New Orleans), recruit a new teacher workforce to the city (e.g., Teach for America and TeachNOLA), and provide information to parents to help them choose schools (New Orleans Parents Guide). The state RSD existed prior to Katrina but had just a handful of staff and had not been designed to carry out its new responsibilities. Hoxby (2000) argues that it would take 10 years to see a radical departure from the Tiebout model reach equilibrium. Given all the changes that occurred, this appears to be a realistic assessment.

To avoid imposing restrictive assumptions of two-period DD and related types of models[13], we instead rely mostly on Granger/event study estimates (Granger, 1969; Autor, 2003; Angrist & Pischke, 2009) as follows:

$$A_{jt} = \gamma_j + \lambda_t + \sum_{\tau=0}^{m} \delta_{-\tau}\left(NOLA \cdot d_{j,-\tau}\right) + \sum_{\tau=1}^{q} \delta_{+\tau}\left(NOLA \cdot d_{j,+\tau}\right) + \varepsilon_{jt} \qquad (2)$$

where λ_t is a vector of year indicators, m is the number of years in the data prior to treatment and q is the number of years after treatment. This implies that $\delta_{-\tau}$ is the adjusted difference in outcomes of the control and treatment groups τ periods before treatment. Since causes must precede effects, these should be insignificantly different from zero and provide a test of parallel trends. If parallel trends holds, then it is reasonable to interpret $\delta_{+\tau}$ as causal effects of the reforms. The estimation of (2) also shows how the effects increase (or decrease) toward the longer-term effects from the estimation of (1).

[13] When there are more than two periods of data, it is common to estimate the following variation of (1) which uses all years of data: $A_{jt} = \gamma_{0j} + \gamma_{1j}t + \delta(NOLA \cdot d_t) + \varepsilon_{jt}$ where t is a continuous time period variable. This specification yields downwardly biased estimates, however, when there are dynamic effects (Pischke, 2005). We show later that the effects of the New Orleans school reforms were dynamic; specifically, that they arose over time through a change in the slope of achievement rather than an intercept shift. Our estimates of equation (2) avoid unnecessary restrictions on effect dynamics.

We use two general strategies to estimate both models: (a) panel analysis using only that portion of the pre-hurricane student population that returned to their pre-hurricane district for at least one year post-hurricane; and (b) pooled cross-sections of student cohorts who were in the same grades pre- and post-hurricane (e.g., comparing achievement for the 2004-05 cohort of 4th graders with the 2011-12 cohort of 4th graders). With the panel approach, we are able to study a fixed group of individuals and thereby account for unobserved differences directly; however, the returning group is a small, non-random subset sample of the original population, which limits statistical power and generalizability. Also, eventually, the pre-treatment students go beyond tested grades, making it impossible to study the longer-term reform effects of primary interest. With the pooled cross sections, the sample is much larger as almost all students who were in New Orleans schools pre- or post-Katrina contribute to the estimation, but we have to rely on observable demographic information to account for population change.

We include the usual parallel trends tests and account for potential endogeneity using a variety of the methods discussed by Bertrand, Duflo, and Mullainathan (2004): graphing the dynamics of the effects (see model (2)), using a triple difference (DDD), adding treatment-specific time trends that vary pre- and post-reform, and looking for an effect prior to intervention (placebo tests).[14] The results are generally robust to these alterations. Since these tests are insufficient with the various potential threats to identification discussed above (population change, strategic behavior, effects of other policies, interim school effects, and trauma/disruption), we take additional steps as well.

[14] BDM (2004) distinguish between triple difference (DDD) and the addition of lagged dependent variables. Since the addition of the lagged dependent variable is on some sense of the addition of a third difference, we refer to this as a DDD.

Aggregation of data to the district-by-year level generally yields conservative standard errors (Bertrand et al., 2004; Angrist & Pischke, 2009). The main alternative is estimation at the student-level with Generalized Estimating Equation (GEE) clustering at the district level (Liang and Zeger, 1986); however, the GEE approach rests on asymptotic assumptions about the number of clusters, which are implausible in this case. Inference is generally only valid with at least 30-50 clusters (Kezdi (2004; Cameron, Gelbach, and Miller, 2008; Angrist & Pishke, 2009). In the analyses here that are restricted to hurricane-affected districts, the number of clusters is generally eight or fewer. As we show below, the results are generally robust to the choice between aggregated and student-level/GEE estimation.[15] While the standard errors from the aggregated regressions are generally most conservative, this is not always true and we always report the largest standard errors.

Data and Matching

Louisiana's systems of testing and data management are similar to a growing number of states. The Louisiana Department of Education (LDOE) provided student-level longitudinally linked data for essentially all public school students in the state. Key variables include student test scores, demographics, grade level, and the schools where students enrolled. Pre- and post-Katrina, students took state standardized tests in grades 3-8. While there is some high school testing data, it is not useful for research.[16]

Table 1 provides descriptive statistics for the test scores for each grade and subject. This shows that New Orleans students were 0.3-0.5 standard deviations below

[15] A third common alternative, the wild bootstrap (Cameron, Gelbach, & Miller, 2008) is infeasible in this case because there is only one treatment cluster.
[16] Louisiana began using End-of-Course (EOC) exams in high school after Katrina though the participation rate changed over time in ways that make those scores difficult to study.

the state average pre-Katrina, which is partly what led the state to institute the reforms. Also, the variance in scores in New Orleans was near the state average before the reforms, but consistently above it in the 2011-12. This may be because of effect heterogeneity that we explore later in the analysis. The table also reinforces the results in Figures 1A-1H showing large increases in test scores after the reforms were put in place.

Our data cover the time period of 2001-02 to 2011-12. This is a convenient end point because most of the major reforms were completed by this point and the system had stabilized in the number of schools and students.[17] In the appendix, we address various data limitations and find no evidence that these affect the results.

Matching

Having a within-state comparison group allows us to account for the differences in the test scale across grades and years, as well as changes in state policy that are unrelated to the New Orleans' school reforms. We narrow the comparison group further to account for trauma/disruption effects that arose in all hurricane-affected districts.[18] If the trauma/disruption effects were the same in New Orleans and other districts, this would eliminate it as a source of bias. That said, there are good reasons to think that New Orleans was harder hit than all but perhaps two districts.[19] Therefore, we view the

[17] Some noteworthy changes that occurred more recently. In 2012, the decentralized enrollment system was replaced with a mostly centralized one where students are assigned by a deferred admission algorithm based on the Nobel-prize willing work of Alvin Roth (Harris, Valant, & Gross, 2015). In 2014, the OPSB and RSD signed an agreement of cooperation and common rules were put in place for special education, expulsion, student enrollment, and facilities.

[18] According to Pane et al. (2006), 81 percent of the displaced students came from Orleans, Jefferson, and Calcasieu Parish. Five additional parishes account for nearly all of the remaining displaced students: St. Tammany, St. Bernard, Plaquemines, Vermilion, and Cameron. Pane et al. (2006) define "displaced" as any student who exited the school system because of the hurricane, as determined by the state government and parishes. We consider all eight parishes to be hurricane-affected in what follows.

[19] Pane et al. (2008) show that New Orleans accounted for more than half the students in the entire state who left their home districts for a long enough period that they enrolled in another Louisiana district or left the state and did not return.

comparison of the statewide and hurricane-affected districts as only a test for whether trauma/disruption played a role.

Panel Matching - Version 1. In the panel analysis, our first matching method (Panel-M1) [20] involves the following specific steps: (a) restrict to hurricane-affected school districts (see above); (b) from those affected districts and schools, drop students who never returned to their pre-Katrina district; and (c) among the returning students, use Mahalanobis matching to identify comparison students with similar composite test score levels in both of the two most recent pre-Katrina years (2004 and 2005), stratifying by year of return. To account for grade repetition, step (c) is further stratified so that students who ever-repeated (never-repeated) a grade pre-Katrina are only matched to other students who ever-repeated (never-repeated) pre-Katrina.[21,22] Step (b) helps ensure that the comparison group is similar to New Orleans in the unobserved factors associated with return to the original district.[23]

We match each New Orleans student to one student in each hurricane-affected district. Matching is with replacement therefore comparison group students are weighted

[20] We have also considered matching based on the degree of hurricane damage experienced by individual schools and neighborhoods, though those data are not available at this time.
[21] In Louisiana, students are retained in grades 4 and 8 if they do not reach the Basic level on one or more tests. (The number of tests for which Basic is required has changed over time.)
[22] Since our main analyses are at the district level, we cannot include bin-by-cohort indicators to account for stratification; however, we do so in the robustness checks where we estimate at the student level. Also, we restrict matches to bins that have at least 10 students.
[23] For example, parents who were unemployed prior to the hurricanes might have evacuated with their children to other districts and found jobs there, reducing the probability of returning to the original district. Since we cannot observe unemployment, and we would expect unemployment to influence student learning, this would introduce bias in the absence of matching. The matched comparison group allows us to account for it directly, to the degree that the factors determining return were the same across districts. There may also have been unobserved factors associated with the neighborhood from which families moved. Residents tend to live near others with similar incomes; if families in some neighborhoods returned sooner than others, then this should mean that the ability to return depended on (unobserved) income, which would affect returnees and non-returnees in similar ways, ceteris paribus.

to reflect the number of students in New Orleans they are matched to, so that the weighted distribution in each district looks much like New Orleans.

Panel Matching – Version 2. The alternative matching method is identical to Panel-M1 except that Panel-M2 also stratifies the on one demographic measure (usually free/reduced price lunch status). This is based on prior evidence that achievement growth varies by student background.

Pooled Matching. For the pooled cross sections, the matching process differs because there are now multiple cohorts of New Orleans students, each of whom requires a comparison group. We can match the pre-reform cohort on pre-reform outcomes, but it is less obvious how to match the post-reform cohorts whose outcomes are endogenous. Our preferred strategy is to match whole *schools* using their pre-reform characteristics and then assume that the unobserved factors affecting achievement in those specific schools were the same after the hurricanes among post-Katrina cohorts in those schools. With this assumption, we can still match the post-reform cohorts but without relying on any post-reform data.

Specifically, for the pooled analysis, we match the post-reform cohorts on pre-reform measures as follows: (a) again, restrict to hurricane-affected districts; (b) identify potential match schools as those that exist in 2002-2005 and in 2012 and have at least 10 students in each tested subject and grade; (c) drop districts that have fewer than four potential school matches; and (d) among these schools, use Mahalanobis matching to identify comparison schools with composite test score levels in 2002.[24] Note that step (b)

[24] We match on 2002 instead of 2004 and 2005 because this yielded a more valid comparison group (i.e., it was more likely to pass the parallel trends test). We considered additional matching methods such as matching on achievement growth instead of levels. These methods often led to non-parallel pre-trends, though the post-trends were unaffected.

applies only to the comparison group; that is, all post-Katrina New Orleans schools count toward the district-level[25] outcomes regardless of whether they existed pre-Katrina.[26]

The differences in matching also highlight an advantage of the pooled identification strategy. One of the threats to identification is that the implementation of NCLB would have increased scores in New Orleans even in the absence of the city's larger reform effort, and done so more than other districts because of the city's disproportionate share of low-performing schools. Since NCLB places pressure on whole schools, matching at the school level, as in the pooled analysis, has some advantages over the panel student-level matching.

Once the panel and pooled matching processes are complete, we also aggregate both the New Orleans and matched comparison group up to the district-by-year level to allow estimation at both the student and district levels. In both cases, and with all the panel and pooled matching methods, we weight comparison group students based on the number of times they are matched to New Orleans students. In the panel analysis, this implies that the weighted number of students is the same in every district because every district is being matched to the same number of New Orleans schools. In the pooled matching, the weighting is similar, except that we match at the school level and therefore we weight based on the number of times each school is used. Since school size varies across districts, this yields some small differences in the weight attached to each district

[25] We use the term "district-level" for simplicity, though recall that New Orleans has two school districts, the RSD and OPSB. When we use the term to refer to New Orleans, we mean all students in public schools located in the city.
[26] Since few non-New Orleans school completely closed as a result of the hurricane, and none of the other districts experienced major reforms, this omits very few schools from the comparison districts prior to the Mahalanobis matching.

in the panel versus the pooled. We also considered using synthetic cohort analysis, though this approach does not have good statistical properties in this situation.[27]

Taken together, these DD/matching strategies at least partly address all of the main threats to validity: The panel DD avoids the issue of population change. The restriction to hurricane-affected districts addresses interim school effects and trauma/disruption. Matching on test scores helps address the threat posed by NCLB (since all low-performing students and schools were pressured to improve scores). Later, we discuss additional methods for addressing population change in the pooled analysis as well a strategic behavior from test-based accountability.

Descriptive Statistics for New Orleans versus Matched Comparison

In addition to the test score information, Table 1 shows that the New Orleans population is extremely disadvantaged with 83-86 percent eligible for free and reduce price lunch (FRPL); almost all the students are racial/ethnic minorities and 93 percent are black. The differences between 2004-05 and 2011-12 also provide a first indication that the demographics of the New Orleans public school population changed relatively little after the hurricane.

Table 2 shows the results of the matching process for Panel-M1. In the panel analysis, the matching process succeeded in finding matched samples of students in hurricane-affected districts that, prior to Katrina, had test score levels similar to New

[27] Synthetic cohort analysis is typically used when there is a single treatment unit (e.g., school district) and there are multiple candidate comparison groups, some of which are more similar to the treatment group at baseline. In this case, we do have a single treatment unit (New Orleans), but almost all the variance is between schools within school districts. More generally, synthetic cohort analysis is not as useful when: (a) there is a common support problem at the level of the policy implementation (i.e., the district); and (b) there are smaller units below the level of policy implementation that are nested. Under these conditions, Mahalanobis matching at the lower-level unit of aggregation is more effective in identifying a reasonable comparison group. In theory, we could do synthetic controls at the district level after doing Mahalanobis matching at the school level, but the Mahalanobis matching removes so much of the variation that this does not appear very useful.

Orleans. Column (5) shows that the panel comparison group is 0.07 standard deviations higher than New Orleans in pre-reform test levels (averaging across subjects and grades). This is far better than the unmatched; columns (1) and (2) show that New Orleans was more than 0.5 standard deviations below the state average. The fact that we can match only at the school level in the pooled analysis clearly makes the match less successful. As a result, the pooled matching method yields a difference between New Orleans and the comparison group of 0.34 standard deviations. We show the parallel trends tests later.

Population Change

One of the main threats to identification in the pooled analysis is that the population may have changed disproportionately in New Orleans relative to the comparison group. As noted earlier, the New Orleans population has similar rates of FRPL participation before and after the reforms (Table 1). However, FRPL is problematic because it cannot capture the difference between students just below the poverty line and those in extreme poverty, and because the FRPL reporting rates depend on how schools administer the FRPL program. We therefore provide additional evidence.

First, Panel A of Table 3 indicates the pre-Katrina 3rd grade characteristics of students who returned to New Orleans, relative to the same figures for the hurricane-affected districts. By 2010, the difference-in-difference (DD) in pre-hurricane achievement of returnees actually favored the comparison districts (by 0.055 standard deviations). The small change in the population is reinforced by the demographic measures; the DD calculations for the percentage of students who are special education, ELL, and FRPL are all less than three percentage points.

Since the above administrative data are somewhat limited (e.g., they only include returnees and the pooled analysis includes all post-Katrina students), we commissioned the U.S. Census Bureau to provide demographics for households with students in public schools, for each district in the state.[28] Panel B of Table 3 shows that some socio-economic measures favor New Orleans and others favor the hurricane-affected districts. For example, median household income dropped by $736 in New Orleans, but increased in the comparison districts by $1,750, for a DD of -$2,486 (2012 dollars).[29] However, the percentage of the population with a BA or higher increased by five percentage points in New Orleans and increased by three percentage points in the comparison group.

To identify the potential influence of these Census-based demographic shifts on student learning, we used data from the USDOE's Early Childhood Longitudinal Study (ECLS) to estimate the partial correlation between achievement levels on each of the demographic measures.[30] With the resulting regression coefficients (shown in Panel C), we then carried out an out-of-sample prediction of the achievement levels/growth change expected from the changes in Census demographic measures.[31] The results are shown in Panel D. The cumulative effect across 4.5 years in the reformed system (our estimate of the "dosage"), averaged across the demographic measures, is 0.012 standard deviations.

[28] The Census could only provide these data for districts with more than 100,000 residents. These are: Calcasieu, Jefferson, and St. Tammany. The results were similar when we looked at the state as a whole.
[29] The absolute decline in socio-economic characteristics in New Orleans is corroborated by Vigdor (2008).
[30] In each regression, the ECLS test score (in levels and growth, respectively) is regressed on one demographic measure and school fixed effects We include only one demographic measure in each regression because the Census demographic changes shown in Panel A do not account for the covariances among them. Also, note that the Census data are for the whole district and 25 percent of the school-age population attends private schools.
[31] We estimate the models separately for achievement levels and achievement growth so that the cumulative predicted effect reflects both. See table notes for details on the different cumulative measures.

and the largest estimate is 0.044 standard deviations.[32] As in Panel A, this suggests a very slight upward bias in the pooled analysis.

Overall, it appears that the elimination of public housing and disproportionate flooding impact on low-income neighborhoods had a minimal effect on the relative demographics of the public school population years after the hurricanes. This is partly because the hurricane affected 80 percent of the city, so that all demographic groups were affected. Also, the number of federal Section 8 public housing vouchers was much larger than the drop in public housing units, so more low-income families, and their children, were apparently able to return than appears at first glance.[33] Finally, note that while extremely poor students were somewhat disproportionately affected, there was also an offsetting reduction in the black middle class. In any event, this suggests that population change is not a major threat to identification in the pooled analysis.

Results

Panel Estimates of Average Treatment Effects

Panel-M1. Table 4 reports results from the panel analysis estimation of average treatment effects (ATEs) based on equation (1) for 4th and 5th graders by year of return. The column (1) sample includes almost all Louisiana students who have data pre- and post-hurricane (without matching)[34]; column (2) includes the entire state matched on test

---

[32] The results in Table 3 are based on reading only and for the entire population. We therefore also re-estimated the Panel C models for low-income ECLS students, which increases the predicted achievement effects, and re-estimated for ECLS math, which reduces the effects, thus the reported effects on reading for the whole population represent a middle ground. We thank Jane Lincove for suggesting these checks.

[33] According to Seicshnaydre and Albright (2015), the number of housing vouchers used increased from 4,763 in 2000 to 8,400 in 2005 (which includes some post-Katrina months) to 17,437 in 2010, for a drop of at least 10,000 units. In contrast, the number of public housing units dropped by about 5,000 units.

[34] We excluded only those students who did not return to their 2005 district for at least one year and students who took alternative assessments. These same exclusions apply to both New Orleans and the comparison districts.

score levels. We follow the same pattern in columns (3) and (4), showing unmatched and matched samples with the hurricane-affected districts, the latter being our preferred specification. The results are reported separately for 2006 and 2007 returnees.[35] Since our test scores end in grade 8, we can follow pre-Katrina 4^{th} (5^{th}) graders only through 2009 (2008). Also, these are cumulative effects where the number of years under the reforms varies directly with the year of return (e.g., the 2009 cumulative effect for 2007 returnees involves three years under the new system).

While about one-third of the 80 estimates are positive and significant, the effects are systematically smaller and generally insignificant in our preferred specification. With a matched comparison group from hurricane-affected districts, the point estimates average about 0.10 standard deviations through 2009 for pre-Katrina 4^{th} graders. The estimates are similar between the state and hurricane-affected districts, but much larger without matching. Since the matching is based on (multiple) test score levels, the sensitivity to matching may mean that NCLB or other statewide policies or a change in the test scale were influencing low-performing schools in other parts of the state.

The effects for 4^{th} graders are noticeably larger for students who returned in 2007, perhaps reflecting either improvement in the school system over time or larger interim school effects for students who returned later. The effects for 5^{th} graders (Panel B) are smaller and include the only two cases in this study where we find negative and significant coefficients. In all cases with the hurricane-affected matched comparison, we

[35] The vast majority of students who returned and who have post-Katrina data in grades 3-8 had returned by 2007. Also, there are very few returnees in other hurricane-affected districts to match with after 2007.

pass a parallel trends test.[36] In 18 of 40 cases with the unmatched results, we reject parallel trends, reinforcing our preference for the matched results.

To leverage the entire panel, and not just two time points in Table 4, we also estimate model (2) (i.e., Granger/event studies). The last year in Figure 2 is the same as Table 4 Panel A for 2012 by construction. There are signs, especially among 4th grade returnees, that the effects in later years emerged from a combination of an initial dip in scores in the first year of return followed by a positive upward trajectory. The negative effects in the first year of return could reflect either low-performance of schools in the early years (followed by improvement) or the especially harsh conditions and trauma of returnees in New Orleans the first few years after the storm.[37]

The specific cause of the initial dip, while difficult to establish, has a significant influence on the interpretation. If the dip is due to trauma and disruption, and that effect fades out in later years, then the reform effect is best estimated by the 2008/2009 estimates, which average to 0.10 standard deviations (statistically significant for 4th graders). The same is true if there is no trauma/disruption effect and the dip is due to a negative initial reform effect.[38]

---

[36] Specifically, we estimate a model that allows the group-specific trends to vary pre- and post-treatment, which serves as both a robustness check and a test for parallel trends. This follows Dee and Jacob (2011) and is sometimes called a comparative interrupted time series (CITS) model. Specifically, we estimate: $A_{ijt} = \gamma_j + +\beta_1 d_t + \beta_2 Years_t + \beta_3(Years_t \times d_t) + \beta_4\left(Years_t \times NOLA_{ij}\right) + \beta_5(d_t \times NOLA_{ij}) + \beta_6(Years_t \times d_t \times NOLA_{ij}) + \varepsilon_{jt}$. The variable $Years_t$ is continuous equal to 0 in 2005, +1 in 2006 and so on; $\gamma_j$is again a vector of district fixed effects. Our estimate of β_4 is therefore a test of the parallel trends assumption; this is ,reported in Table 4 using two pre-treatment years. Figure 2B shows that parallel trends often do not hold if we use additional pre-reform years (for pre-Katrina 5th graders who returned). We also carried out placebo tests and these yielded similar results (available upon request).

[37] Stratification based on year of return reduces the quality of the match on test levels. Therefore, as a robustness check, we re-estimated by: (a) matching on test scores and year of return (which reduces extremely poor matches on test levels while sacrificing similarity on year of return). The results were quite similar (available upon request).

[38] The results for the 2005 5th graders are in the appendix. They display the same general upward pattern, though it is flatter. Also, the matching process in that case does not satisfy the parallel trends assumption and there are only a maximum of two post-reform years to consider.

Panel-M2. The panel results are generally similar to Panel-M1 when we also match on income (FRPL), as in Panel A of Figure 3. However, Panel B shows that when we implement panel-M2 matching on race the estimated effects are much larger. Since there is no obviously preferred matching method, we establish bounds later by using the average of the various methods.

Overall, the vast majority of coefficients in Table 4 are positive (and almost half of those are precisely estimated), and they are consistently larger for students who have more post-Katrina years to experience reform effects. Also, in 16 of the 20 cohort-by-subject figures we see a positive trajectory over time in the point estimates.[39] A key disadvantage of the panel analysis, however, is that it stops in 2009 and prevents us from testing whether the upward trajectory continues. This might be considered a short span of time to implement an entirely new type of schooling system as well as recruit, select, and create new schools. In 2009, most schools were still being operated directly by the RSD and the majority of teachers were still those from the pre-Katrina period. Only three schools had been closed or turned over to other operators in this time frame, compared with 45 schools between 2008 and 2015. Also, even if the system had reached equilibrium, students would have had fewer years to experience it (a maximum of 3.5 grades for the spring 2006 returnees). Finally, there are some indications here that trauma/disruption effects may have been larger for New Orleans students and pulled down the measured effects in the short term. The limitation of this short time span is addressed by the analysis that follows.

[39] These changes over time are not statistically significant.

Pooled Estimates of Average Treatment Effects

We estimate equation (1) comparing different cohorts of students who took tests in the same grades in New Orleans before and after the hurricanes. As in Table 4, we start by reporting two-period estimates, one pre- and one post-reform, but in this case the latter period is 2011-12, three years later than the panel. Again, these are cumulative effects and students enrolled in New Orleans taking the test in 2012 averaged 4.39 years under the reformed system.[40]

These pooled results, shown in Table 5, are positive for every specification and statistically significant in 91 of 96 cases. Averaging across grades, and focusing just on the hurricane-affected matched sample, the estimates are all positive and statistically significant, in the range of 0.30-0.48 standard deviations across subjects.[41] As in the panel analysis, Figure 4 also suggests that the positive effects are the result of steady improvement leading up to 2011-12. Considering both Table 5 and Figure 4, we can see that the estimates usually pass a parallel trends test, but not always.[42]

Since one of the main threats to identification in the pooled analysis is the change in population, recall that our various estimates in Table 3 suggest very small population changes. Also, the trends in achievement effects are inconsistent with those of population change: we find evidence of an initial upward spike in socioeconomic status in New Orleans right after the hurricanes, which dissipated in the ensuing few years. If

[40] We include in this "dosage" calculation the number of years in both tested (3-8) and non-tested grades (K-2) since most of the reform policies (with the exception of test-based accountability) applied to all grades.
[41] This range is from the "combined" row, which includes all grades. There is a wider range if the results are broken down further by grade and subject.
[42] Given that this method sometimes failed on parallel trends, we also varied the matching method, e.g., matching on trends versus levels and using different combinations of years; these variations performed more poorly with regard to the parallel trends assumption, though the post treatment patterns were nearly identical.

population change were the driving force behind the effects, then we would have expected a large initial achievement effect followed by a flat or declining effect trend. This is almost the opposite of the actual trend, reinforcing the idea that population change does not bias the pooled estimates.

There are also no signs that that the pooled effects were driven by interim schools. Table 5 shows results for 3rd graders in 2012 and these students would not have been of school age until 2009, after the vast majority of evacuees had returned. More generally, compared with other grades, few of the 2012 3rd graders were ever in non-New Orleans public schools. Yet, we see no signs that the effects are smaller for this group.

Robustness Checks and Additional Identification Strategies

Estimation at the Individual Level. We report most of the results aggregated to the district level to obtain conservative standard errors. A disadvantage of this approach is that we cannot add student-level covariates, such as demographics and whether students had been retained in grade, or a vector of bin indicators to address stratified matching.[43] We therefore re-estimated the models at the student level adding these covariates, though this had only a minimal influence on the results. In both the panel and pooled methods, estimation at the student level, with standard errors clustered at the district level, has a minimal effect on precision (see the appendix). The effects are qualitatively similar when switching the dependent variable to achievement growth (a form of triple difference).[44] We also find no evidence of bias from missing data.[45]

---

[43] These covariates could not be included in the main models because these are estimated at the district level of aggregation. Identification of these parameters at the district-level is based on changes in district-level demographics over time, which are extremely small and have little variance across districts, resulting in implausible parameter estimates.

[44] Specifically, we estimated the first difference becomes 3rd-to-4th grade growth for the 2010-11 cohort of 3rd graders minus 3rd-to-4th grade cohort in the 2003-04 cohort of 3rd graders. Thus, there are two

Alternative Identification Strategy: District Switchers. We attempted a third identification strategy using only students who switch into or out New Orleans ("in-switchers" and "out-switchers," respectively) and who remain in their new districts within the pre- or post-reform periods. In the simplest model, we essentially take the one-year difference in achievement for individual students before and after the switch (*within* the pre-reform and post-reform periods) and compare this growth before and after the reforms. We also estimate a version of the model that accounts for changes in statewide trends in cross-district mobility.[46]

The identifying assumption of the first simpler model is that the unobserved factors affecting both district of enrollment and achievement are constant over time. In the second model, the assumption is weaker: that the unobserved factors associated with cross-district mobility follow the same trend in New Orleans and the rest of the state. If the switcher strategy is identified, the expected value of the in-switcher effect would be

dimensions of changes over time in this case: within student over time and across cohorts over time. This can provide additional protection against violations of the parallel trends assumption as in a typical triple difference (DDD) models, although our preferred DD method described in the main text seems to satisfy the parallel trends assumption. Nevertheless, while the DDD increases measurement error in the dependent variable, two of the four DDD estimates are statistically significant (science and social studies) and the average point estimate is 0.07 standard deviations in annual growth. These are naturally smaller than the cumulative estimates reported in the main text.

[45] To test whether missing data might explain some of the results, we created a variable for whether a test score is missing and then used this as the dependent variable in model (1). The results suggest there was a slight increase in missingness in 2007, but no differences in subsequent years. Since the matching was based on (observed) test scores, this analysis is necessarily unmatched. Also, this analysis is only done for students who show up enrolled in a school. Other students may be missing from the data entirely because they were not enrolled anywhere.

[46] Specifically, the model for the switcher strategy is: $A_{it} = \lambda A_{i,t-1} + \theta_g + \beta_1 d_t + \beta_2 InSwitch_{it} + \beta_3(InSwitch_{it} \times d_t) + \varepsilon_{jt}$. Our Model 1 (Switcher-M1) includes only lagged achievement of student *i* in time *t* ($A_{ij,t-1}$), a vector of grade fixed effects (θ_g), and an indicator for the post-Katrina period (d_t). In this model, we are interested in $\beta_1$which simply comparing achievement growth from switches that occur before and after the reforms. In Switcher-M2, we also account for the possibility that the types of students who switch changed over time across the entire state. This involves adding $InSwitch_{it}$ as an indicator for whether the switch was specifically into New Orleans ($InSwitch_{it} = 0$ for cross-district switches where New Orleans is neither the sender nor the receiver). In this second model, we are primarily interested in β_3. We then carry out the same estimation replacing $InSwitch_{it}$ with $OutSwitch_{it}$. Unlike the pooled and panel strategies, there is no matching involved.

of the same magnitude as the out-switcher effect, but with the opposite sign. This switching strategy has also initial intuitive appeal because it relies solely on individuals who experienced both New Orleans and non-New Orleans schools and then compares their experiences under the two alternative policy regimes (across cohorts).

To compare the point estimates of the switcher strategy with the others, we re-estimated (1) with achievement gains as the dependent variable. These results are similar to the *differences* in magnitudes between the in-switcher and out-switcher coefficients (i.e., 0.05-0.10 standard deviations) in the pooled analysis. That said, there are two reasons to downplay the switcher results. First, the identifying assumptions do not appear to hold; the in-switcher coefficients are not of equal and opposite sign to the out-switcher coefficients. Also, this strategy requires restricting the sample to roughly 10 percent of New Orleans students, a very small and possibly unusual sample.[47]

ATE Bounds

The effects vary somewhat by method, and some only cover the first few years post-reform, therefore we establish bounds for the longer-term effects using various extreme assumptions. Our first lower bound estimate starts with the pooled analysis and makes an adjustment for estimated bias; it assumes that: (a) the average of Panel-M1 and Panel-M2 is unbiased; and (b) the pooled bias is fixed in magnitude over time. The difference between the panel and pooled serves as an estimate of the bias and we use that to adjust the pooled estimates downward.

The second lower bound is based on linear projection of early panel results into the future; it assumes that: (a) the average of Panel-M1 and Panel-M2 is unbiased; and

[47] We thank Andrew McEachin for suggesting this approach.

(b) the effects continued on the same path after 2009 (consistent with every panel and pooled analysis). In this case, we ignore the pooled results entirely. The two lower bounds yield quite similar effects, +0.20 and +0.23 standard deviations, respectively.

The upper bound is based strictly on the pooled estimates and assumes they are unbiased. This is not implausible given the apparently minimal changes in demographics, the school-level focus of test-based accountability (which implies school-level matching), and the fact that this ignores any lingering effects of trauma/disruption. This yields an upper bound in 2012 of +0.40 standard deviations, a figure that happens to coincide with the total improvement relative to the state reported in Figures 1A-1H.[48] Also, note that we are aware of no alternative theory that could easily explain the upward trend in scores.[49]

The above calculations are summarized in Table 7. We also provide a cost-benefit analysis based on the prior work of Krueger (2003) and Harris (2009), using estimates of the labor market returns to cognitive skill measured by test scores. These results suggest that even the lower bound effects are six times larger than the break-even effect size and larger than commonly discussed policy alternatives, such as reducing class size and increase access to pre-kindergarten education.

Effect Heterogeneity

One of the most common critiques is that the New Orleans school reforms have been inequitable and even harmful to disadvantaged students. Numerous media reports and lawsuits have alleged denied admission, disproportionate suspensions and

[48] An additional assumption is that the effects in elementary and middle school do not extend to high school, which we cannot observe.

[49] One possibility is that, if the state had simply continued the less aggressive pre-Katrina role of the RSD, that this would have generated similar effects. However, note that: (a) there is no strong evidence of this in the pre-trends; and (b) the RSD role is arguably part of the reform package. Since this also affects the control group, this may be generating a downward bias in our effect estimates.

expulsions, and insufficient services among certain disadvantaged students under the city's reforms (*P.B. v. Pastorek*, 2010; Jabbar, 2015).

We carried out the same basic estimation methods as above, but separately by FRPL, race/ethnicity, and special education.[50] The earlier matching process was modified to add stratification by subgroup.[51] In both the panel and pooled cases, we also carried out many of the same robustness and bias checks for each subgroup. In general, the subgroup analyses pass the tests and are robust, though we note a few exceptions below.

The Granger/event study results for the 2007 returnees are shown in Figure 4 for the panel analysis and Figure 5 for the pooled. We include only math and language arts for simplicity, though the results are similar for science and social studies. The effects are positive and significantly different from zero for every subgroup except when the matching stratification is based on special education.[52]

The confidence intervals test whether the effect for each subgroup is different from zero. In only a few cases are the differences in effects between subgroups statistically different from one another and this occurs only in effects during the first year that students returned.[53] In the panel analysis, black and FRPL students have lower initial

[50] We omit English Language Learners (ELL) because there are so few students in this subgroup in New Orleans, and even fewer in the potential comparison districts.
[51] Attempting to match on all of the demographic measures simultaneously led to extremely poor matches on test scores.) In the pooled subgroup matching, we also restricted the comparison group to schools that had at least 10 students in the given subgroup (e.g., 10 in FRPL and 10 non-FRPL); also, we matched on the test scores of each pair of subgroups simultaneously; for example, for each New Orleans school, we looked for a comparison school where FRPL students had similar test scores to the FRPL students in the New Orleans school and where the non-FRPL students in the potential comparison also had scores similar to the non-FRPL students in the New Orleans school.
[52] These figures also show that the effects usually pass a parallel trends test for race and special education, though not always under FRPL. Separately, we also compared New Orleans and the comparison subgroups on test levels. As with the ATEs, the test levels match well in the panel analysis and poorly in the pooled; specifically, New Orleans white students' pre-Katrina scores are considerably above their comparison group means, while New Orleans' black students are below the comparison group.
[53] Even in the few cases where the subgroup effects do seem statistically different from one another, there are many subgroups comparisons and some differences are bound to emerge by chance alone.

effects, but this is followed by similar upward trajectories. This is also true for blacks in the pooled analysis, but not FRPL students. For FRPL students, the differences between the pooled and panel results may be due to the fact that almost all New Orleans' public school students could be considered "homeless" when they first returned and this automatically made them eligible for FRPL.[54]

There are two possible interpretations of why the initial dip in scores post-reform, which we first identified in the analysis of ATEs, might only apply to disadvantaged students. First, black, low-income, and less educated families, who make up the vast majority of New Orleans' public school population (see Table 1), were harder hit by the hurricane in terms of health (Sastry & Gregory, 2013), housing (Elliott & Pais, 2006), and employment (Fussel, 2015; Sharkey, 2007).[55] Perhaps not coincidentally, these same families also experienced worse initial psychological effects (Brown et al. 2011; DeSalvo, et al., 2007; Elliott & Pais, 2006).[56] We also considered whether the dip for

[54] For FRPL purposes, a student is considered homeless if "s/he is identified as lacking a fixed, regular and adequate nighttime residence by the LEA homeless liaison, or by the director of a homeless shelter" (USDA, 2014). Many students were living with relatives or in homes that were still heavily damaged. Thus, even some students who are otherwise socio-economically advantaged could be considered homeless and eligible for FRPL. Since FRPL students are only compared with other FRPL students, this likely led to what appear to be large achievement effects at first and then smaller effects. Further, this pattern would not appear in the panel analysis because FRPL eligibility in that case is based entirely on pre-Katrina FRPL eligibility. We thank Lindsay Bell Weixler for pointing out this issue with the FRPL homeless designation.

[55] According to Elliot and Pais (2006), black and low-income residents were, other things equal, less likely to evacuate prior to the storm and live in a rental or shelter (versus a home they own) in the immediate aftermath. Among adults who were employed prior to Katrina, blacks and low-income people were less likely to be employed after the hurricanes. Blacks also reported more stress with regard to their current circumstances and future prospects. In their study of the probability of return to New Orleans, Paxson and Rouse (2008) find that blacks and families with children were less likely to return, perhaps in part because the rental housing stock declined even more than owner-occupied housing (Vigdor, 2008). Finally, Sharkey (2007) finds a positive correlation between the number of dead bodies found and the neighborhood percentage of residents who were black.

[56] The DeSalvo et al. results are based on a sample of the faculty and staff of Tulane University. They did not find differences by race, but did by income and education levels. Interestingly, while the initial effects on less advantaged families seem to have been worse, there is some evidence that they also seemed to recover faster (McLaughlin et al. 2011).

disadvantaged students might have been due to disproportionately low-performing interim schools, but our results are inconsistent with that theory.[57]

An alternative theory is that New Orleans schools after reforms were less effective in helping disadvantaged students, and they continued to be less effective over time. This theory is consistent with the lawsuits and anecdotal evidence about how the schools operated just after the reforms were put in place. It is difficult to distinguish between the trauma/disruption and system effectiveness hypotheses, however. As noted earlier, because students took tests at the *end* of their first year after returning, so their scores in the first year of return reflect both trauma/disruption and the effectiveness of the schools that first year.

The special education results are highly inconsistent between the two identification strategies, and differ from the other subgroups. The panel results in Figure 4 suggest that special education and non-special education students experienced almost identical, null reform effects, while the pooled analysis shows uneven trajectories and much more positive effects for both groups in the longer run.

The unusual patterns with special education are likely due to two potential biases: endogeneity in the probability of being labeled special education and the probability of taking an alternative assessment conditional on being in special education. School personnel are the ones who place students into special education programs. The fact that we matched students stratifying on their pre-Katrina characteristics helps address potential endogeneity issues in the panel analysis only. The general problems with

[57] Specifically, we calculated the mean 2005 test score levels of the interim schools attended by evacuees in 2006. Using the simple DD model in equation (1), it appears that the racial/income gaps in school quality among New Orleans students dropped when they switched to interim schools, i.e., disadvantaged students experienced large gains on this crude measure of school quality.

matching at the school level in the pooled analysis also apply here, except they are worse because we are still identifying effects from students labeled special education post-Katrina.[58] For this reason, we have more confidence in the panel analysis, which avoids this endogeneity problem. However, both the panel and the pooled suffer from the fact that there is a 10-percentage point effect of the reforms on the probability that special education students took alternative assessments.[59] Only special education students taking the LEAP tests are included and we are continuing to examine ways of adding in students who took alternative assessments.

For all the various subgroup categories, we carried out the same set of checks as with the ATEs and the results are highly robust. We were particularly focused on grade repetition since students in the various disadvantaged groups are more prone to repeat grades, especially in New Orleans, but including grade repetition as a covariate has a minimal impact on the results.[60]

While these results are exploratory and there are some inconsistencies, three clear patterns emerge. First, there is no evidence that any disadvantaged group was worse off academically as a result of the reforms. In the last year of all the figures, for all the subgroups, the effects are positive and often large and statistically significant. Second, with one exception, the disadvantaged groups always see a smaller effect than the advantaged groups early in the reforms. Unfortunately, we cannot determine to what

[58] We estimated a version of model (1) with the dependent variable as the probability of being assigned to special education. In this case, there was a 2-4 percentage point drop in the probability of being assigned to special education, the same pattern we see in the trajectory of effects on achievement of special education students in the pooled analysis.
[59] This is based on the estimation of model (1) with the probability of taking the alternative assessment, among special education students only.
[60] Grade repetition is a greater potential threat to identification in the pooled analysis because we could not successfully match at the individual level. In the panel analysis, we stratified the matching on both grade repetition and subgroup status.

degree this is caused by the reforms themselves or how effective the schools were with disadvantaged students. Overall, it would be hard to say that the New Orleans' reforms were more inequitable than the prior system, especially in more recent years, though the pre-reform system is arguably a low bar.

The third pattern has more to do with the interpretation of the ATEs. Recall that the ATE matching process did not involve stratification on student demographics because this reduces the quality of the match on student test scores. However, since the effect heterogeneity analyses require stratification on student demographics, we can view the weighted average of the subgroup effects as an alternative method to panel estimation of the ATEs. In general, the weighted average of the subgroup effects are considerably larger than the equivalent ATE reported in Table 4. This is especially noteworthy with the panel results (in 2009) where the implied ATE is now about 0.15 standard deviations and marginally significant. This reinforces the notion that the panel ATEs are not unbiased.

Additional Evidence

Strategic behavior from test-based accountability remains perhaps the most plausible remaining source of bias because it is hard to test for strategic behavior in one measure without a separate low-stakes measure to compare with. Such an "audit" test does not exist in Louisiana. Instead, we leverage the fact that the stakes are somewhat higher with math and ELA. Not only are these scores more commonly reported in newspapers, but in some of the years and grades under consideration, they also comprised a smaller portion of the school performance score used to grade, and potentially shut down, low-performing schools.

One of the most consistent findings in this study is that the results do not vary systematically with the stakes. In both panel methods, and the pooled analysis, the average effects are quite similar when we average math with ELA and science with social studies. As further evidence, we considered other outcomes that are even lower stakes than social studies and science: the Louisiana Department of Education (LDOE) reports that high school graduation and on-time college entry (conditional on high school graduation) each improved by 8-10 percentage points in New Orleans compared with the state between 2004 and 2014 (LDOE, 2015). The fact that college entry is increasing at the same time as high school graduation is noteworthy since we might expect the marginal high school graduate to be less likely to attend college.[61] This is also consistent with recent evidence that positive effects on high-stakes tests are associated with positive effects on a range of long-term outcomes (Deming, Cohodes, Jennings, & Jencks, 2015).

Given these large changes in both achievement and other student outcomes, we would also expect to see other changes in practices and other "leading indicators" within the school system, which we are exploring in a number of other studies: (a) with attendance zones eliminated, families became more active choosers with students rarely attending the school closest to home under the reformed school choice system (Harris & Larsen, 2015); (b) schools are differentiated in the types of programs they provide, making good matches with family preferences more likely (Arce-Trigatti, Lincove, Harris & Jabbar 2015); (c) the state RSD is opening and closing based on demonstrated evidence of success in generating student achievement (Ruble & Harris, 2015); and (d)

[61] It is possible that both measures are biased. In particular, there is some evidence that RSD schools are labeling too many students as out-of-state transfers. If some of these students are actually dropouts, this would inflate both the high school graduation rate and the college entry rate. We are in the process of obtaining the exit codes and college entry data to carry out our own analysis, akin to the test score analysis.

the teacher workforce changed significantly and in ways plausibly consistent with achievement growth (Barnett & Harris, 2015).

There are also some places where we might have expected negative consequences that did not emerge. Voluntary student mobility has remained largely unchanged in New Orleans relative to the state as a whole (Maroulis, Santillano, Jabbar, & Harris, 2015), perhaps because the choice system leads to better initial matching of students to schools, reducing the need to switch schools (Harris, Valant, & Gross 2015). Some have also worried that the reforms would increase racial and income-based segregation across schools, though there is limited evidence of this either (Barrett, Weixler, Zimmer, & Harris, 2015). With many apparently positive changes in the operation of the school system, and relatively few negative changes, it is perhaps not surprising that student outcomes improved so much, on average and for disadvantaged students.

Conclusion

For more than half a century, the U.S. public school system has followed a fairly uniform model across the country in which schools are controlled by locally elected school boards that set attendance zones for students and negotiate contracts with teacher unions. Opponents, dating back at least to Friedman (1962), argue that this system provides limited incentives for improvement (Chubb & Moe, 1990; Hill, Pierce, & Guthrie, 1992), resulting instagnant test scores and declining economic competitiveness (National Commision on Educational Excellence, 1983; Hanusehek & Woessman, 2010).

These arguments of reformers led to the test-based and market-based accountability movements of the past two decades (Harris & Witte, 2011). Modestly implemented and modestly successful to date, a key lingering question has been whether these general strategies might be more or less effective if they were designed and implemented with greater intensity. Testing this theory has been difficult, however, because there have been almost no counter-examples to the traditional system of public education in the U.S. Even when partial examples have emerged (e.g., in Washington, DC and New York City), identification problems have been daunting. Systemic school changes are difficult to study, since they tend to roll out gradually over time and are mixed in with other policy and demographic changes. In New Orleans, in contrast, the policy shift was sudden and intense.

We find that that the package of market- and test-based accountability policies put in place after Hurricane Katrina increased student achievement by 0.2-0.4 standard deviations. This translates to somewhere between 50-100% of the total improvement New Orleans experienced relative to the state. There are some signs that the implementation of NCLB would have generated some of these same effects in the absence of the New Orleans reforms, but note that NCLB is based on similar principles.

The distinctive methodological challenges to studying reforms coming in the wake of a natural disaster do not appear to introduce much bias. None of our three types of analysis suggests that population change could explain more than 10 percent of our upper bound reform estimates. The net effects of interim schools and trauma/disruption also seem very small. The worst-case scenario appears to be an upward bias of no more than 10 percent of the point estimates, and it appears equally likely that the bias from these factors is actually downward.

Nevertheless, the fact that the reforms seem to have been beneficial on average and for key subgroups in New Orleans does not mean these benefits would extend to other cities. In general, external validity considerations rest on the types of participants served, the intensity and quality of policy implementation, and the basis of comparison. In this case, the participants were entirely black and low-income students with test scores that were extremely low, even by urban district standards. The New Orleans reforms were also implemented with an unusual, and perhaps unusually large and high-quality, supply of educators. There was a national out-pouring of support from across the nation. People flocked to the city to help rebuild and many stayed. The city also became an epicenter for school reform and a magnet for ambitious, talented, young educators from around the country.

While the reforms were implemented in an entire school district, taking the policy to a larger scale, such as a whole state, could prove more challenging. Teacher quality again comes into play because the supply of educators from Teach for America and other more elite alternative preparation programs is limited. New Orleans is also a relatively small district, especially after Katrina, and requires relatively few teachers. Taking New Orleans-style reforms to larger districts, or simply more districts, would require larges shifts in teacher supply.

Finally, the basis of comparison in this difference-in-difference analysis is a pre-Katrina school system that, by just about any measure, was failing badly. Corruption, mismanagement, and rapid turnover of superintendents resulted in extremely poor student outcomes (Council of Great City Schools, 2001; Buerger & Harris, 2015, Cowen Institute, 2015; Perry, Harris, & Buerger, 2015). There may be diminishing returns to system reform and districts that have pursued other types of reform might see smaller effects from New Orleans-style policies as a result. Put differently, New Orleans had nowhere to go but up.

While the generalizability of the findings are, as always, a bit unclear, there is much to be learned here. More than a decade ago, Hoxby (2000) speculated on how hard it might be to ever observe the effects of a massive reform in a U.S. school system, yet the conditions she described are quite similar to what we see in New Orleans.[62] The successes documented here force educators and policymakers to question assumptions about how an education system can and should be designed and operated. It shows that, at least under certain circumstances, intensive system-wide school reform, based on principles of accountability and school autonomy, have the potential to produce large effects on student learning. The question now is whether such large gains can be achieved at scale in other cities, through these or other means, without a tragedy like Hurricane Katrina.

[62] Hoxby (2000) writes that the "Tiebout process . . . is still the most powerful force in American schooling. It will be years before any reform could have the pervasive effects that Tiebout choice has had on American schools. Moreover, the short-term effects of reforms [would be] misleading because … the supply response to a reform--the entry or expansion of successful schools and the shrinking or exit of unsuccessful schools--may take a decade or more to fully evince itself."

References

America Next (2015). *K-12 Education Reform: A Roadmap*. Downloaded April 27, 2015 from: http://americanext.org/wp-content/uploads/2015/02/America-Next-K-12-Education-Reform.pdf.

Abdulkadiroğlu, A., Angrist, J.D., Hull, P.D., & Pathak, P.A. (2014). Charters Without Lotteries: Testing Takeovers in New Orleans and Boston. *NBER Working Paper No. 20792*. Cambridge, MA; National Bureau of Economic Research.

Angrist, J.D., Abdulkadiroglu, A., Dynarski, S., Kane, T.J., & Pathak, P. (2011a) Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots. *The Quarterly Journal of Economics*.

Angrist, J.D., Cohodes, S.R., Dynarski, S.M., Pathak, P.A., & Walters, C.R. (forthcoming). Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry and Choice. *Journal of Labor Economics*.

Angrist, J.D., Dynarski, S.M., Kane, T.J., Pathak, P.A., & Walters, C.R. (2010). Inputs and Impacts in Charter Schools: KIPP Lynn?, *American Economic Review (Papers and Proceedings)* 100:1-5.

Angrist, J.D., Pathak, P., & Walters, C.R. (2011b). Explaining Charter School Effectiveness. Working Paper 17332. Cambridge, MA: National Bureau of Economic Research.

Angrist, J. & Pischke J-S. (2009). *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Arce-Trigatti, P., Harris, D., Lincove, J., & Jabbar, H. (2015). Many Options in New Orleans Public Schools. *Education Next* 15(4), 25-33.

Athey, S. & Imbens, G. (2003). Identification and Inference in Nonlinear Difference-in-Differences Models. *Econometrica* 74(2): 431-497.

Barrett, N. & Harris, D. (2015). *Significant Changes in the Teacher Workforce*. New Orleans, LA: Tulane University, Education Research Alliance for New Orleans.

Belfield, C.R. & Levin, H.M. (2003). The effects of competition on educational outcomes: a review of US evidence. *Review of Educational Research* 72(2): 279-341.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1), 249-275.

Bewley, T., 1981. A critique of Tiebout's Theory. Econometrica 49(3), 713-740.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system, American Educational Research Journal. 42(2), 231-268.

Brown, T.H., Mellman, T.A., Alfano, C.A., & Weems, D.F. (2011). Sleep Fears, Sleep Disturbance, and PTSD Symptoms in Minority Youth Exposed to Hurricane Katrina. *Journal of Traumatic Stress* 24(5), 575–580.

Buerger, C. (2015). Orleans Parish Revenues and Expenditures. Presentation at *The Urban Education Future? Lessons from New Orleans 10 Years After Hurricane Katrina*. June 18-20, New Orleans, LA.

Buerger, C., & Harris, D., (2015). How Can Decentralized Systems Solve System-Level Problems? An Analysis of Market-Driven New Orleans School Reforms. *American Behavioral Scientist*. 59(10) 1246–1262.

Card, D. & Krueger, A. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review* 84(4), 772-793.

Carnoy, M. & Loeb, S. (2003). Does external accountability affect student outcomes? A cross-state analysis. Educational Evaluation and Policy Analysis, 24(4), 305-331.

Center for Research on Education Outcomes (2013a). National Charter School Study. Palo Alto, CA: Stanford University.

Center for Research on Education Outcomes (2013b). Charter School Performance in Louisiana. Palo Alto, CA: Stanford University.

Chubb, J.E., & Moe, T.M. (1990). Politics, Markets, and Schools. Washington, DC: Brookings Institution.

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The Effectiveness of Secondary Math Teachers from Teach For America and the Teaching Fellows Programs*. NCEE 2013-4015. National Center for Education Evaluation and Regional Assistance. Retrieved from http://eric.ed.gov/?id=ED544171

Council of Great City Schools (2001). *Rebuilding Human Resources in New Orleans Public Schools*. Washington, DC.

Cowen Institute for Public Education Initiatives (2013). State of Public Education in New Orleans. New Orleans, LA; Tulane University.

Cowen Institute for Public Education Initiatives (2015). State of Public Education in New Orleans. New Orleans, LA; Tulane University.

The Data Center (2014). Who lives in New Orleans and metro parishes now? Downloaded April 25, 2015 from: https://gnocdc.s3.amazonaws.com/reports/TheDataCenter_WhoLivesInNewOrleansAndMetroParishesNow.pdf.

Dee, T. & Jacob, B. (2011). The impact of no Child Left Behind on student achievement, *Journal of Policy Analysis and Management* 30(3), 418-446.

DeSalvo, K. B., Hyre, A.D., Ompad, D.C., Menke, A., Tynes, L.L., & Muntner, P. (2007). Symptoms of Posttraumatic Stress Disorder in a New Orleans Workforce Following Hurricane Katrina. *Journal of Urban Health* 84(2), 142-152.

Elliott, J.R. & Pais, J. (2006), Race, class, and Hurricane Katrina: Social differences in human responses to disaster. *Social Science Research 35*, 295–321

Epple, D., Romano, R., & Zimmer, R. (2015). Charter Schools: *A Survey of Research on Their Charaacteristics and Effectiveness*. NBER Working Paper 21256. Cambridge, MA: National Bureau of Economic Research.

Evers, W. (2014) Implementing Standards and Testing, In Chester E. Finn and Richard Sousa. *What Lies Ahead for America's Children and Their Schools*. Stanford, CA: Hoover Institution Press.

Figlio, D. (2006). Testing, crime and punishment. Journal of Public Economics. 2006, 90(4), 837-851.

Figlio, D.N. & Lucas, M.E. (2004). "What's in a grade? School report cards and the housing market," American Economic Review, 94(3), 591-604.

Friedman, M. (1962) "The role of government in education" in *Capitalism and Freedom*, Chicago: University of Chicago Press.

Fryer, R.G. (2014). Injecting Charter School Best Practices into Traditional PublicSchools: Evidence From Field Experiments. *Quarterly Journal of Economics*. 129(3):1355-1407.

Gill, B. & Booker, K. (2008) School Competition and Student Outcomes. In Helen F. Ladd and Edward B. Fiske (Eds) *Handbook of Research in Education Finance and Policy* (pp.183-202). New York: Routledge.

Government Accounting Office (2007). *No Child Left Behind Act: Education Should Clarify Guidance and Address Potential Compliance Issues for Schools in Corrective Action and Restructuring Status*. GAO-07-1035. Washington, D.C.

Gill, B., Zimmer, R., Christman, J, Blanc, S. (2006) State Takeover, School Restructuring, Private Management, and Student Achievement in Philadelphia. Santa Moica, CA: RAND Corporation.

Goldin, C. & Katz, L. (2008). The Race Between Education and Technology. Cambridge: Harvard University Press.

Groen, J. & Polivka, A. (2008). The Effect of Hurricane Katrina on the Labor Market Outcomes of Evacuees. *American Economic Review*, 98(2): 43–48.

Hanushek, E.A. (1996). A More Complete Picture of School Resource Policies. *Review of Educational Research* 66(3): 397-409

Hanushek, E.A. & Woessman, L. (2010). The High Cost of Low Educational Performance: The Long Run Impact of Improving PISA Outcomes. Paris: Organization for Economic Development and Cooperation.

Harris, D. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis*. 31(1): 3-29.

Harris, D. & Herrington, C. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, 112(2): 209-238.

Harris, D. & Larsen, M. (2015). What Schools Do Families Parents Want (and Why)? Academic Quality, Extracurricular Activities, and Indirect Costs in New Orleans Post-Katrina School Reforms. New Orleans, LA: Education Research Alliance for New Orleans, Tulane University.

Harris, D., Valant, J., & Gross, B. (2015). The New Orleans OneApp. *Education Next* 15(4), 17-22.

Harris, D. & Witte, J. (2011). The market for education. In D.E. Mitchell, R. Crowson, and D. Shipps (Ed.), Shaping Education Policy: Power and Process. New York: Routledge.

Hill, P. & Campbell, C. (2011). *Growing Number of Districts Seek Bold Change With Portfolio Strategy*. University of Washington: Center for Reinventing Public Education.

Hill, P. & Lake, R. (2004). *Charter Schools and Accountability in Public Education*. Washington, DC: Brookings Institution Press.

Hill, P. L. Pierce, J. Guthrie (1997) *Reinventing Public Education: How Contracting Can Transform America's Schools*. Chicago: University of Chicago Press.

Hoxby, C.M. (1996). How Teachers' Unions Affect Education Production, Quarterly Journal of Economics, 111(3), 671-718.

Hoxby, C. (2000). Competition among Public Schools Benefit Students and Taxpayers? American Economic Review 90(5), 1209-1238.

Hoxby, C. (2002). School Choice and School Productivity (Or Could School Choice be a Tide that Lifts All Boats). NBER Working Paper 8873. Cambridge, MA: National Bureau of Economic Research.

Jacob, B.A. (2005), "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools", Journal of Public Economics, 89: 761-796.

Kollman, K., Miller, J.H., & Page, S.E. (1997). Political institutions and sorting in a Tiebout model. American Economic Review 87, 977-92.

Koretz, D. (2009). Measuring Up: What Educational Testing Really Tells Us. Cambridge, MA; Cambridge University Press.

Krueger, A. B. (2003). Economic considerations and class size. Economic Journal, 113, 34–63.

Krueger, A.B. & Zhu, P. (2004). Another Look at the New York City Voucher Experiment. *American Behavioral Scientist* 47(5): 658-98

Lee, J. (2008). Test-Driven External Accountability Effective? Synthesizing the Evidence from Cross-State Causal-Comparative and Correlational Studies. *Review of Educational Research* 78(3), 608-644.

Liang, K-Y. & Zeger, S.L. (1986. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 73(1): 13-22.

Louisiana Department of Education (2005). Louisiana Department of Education. (2015). High school performance. Retrieved from http://www.louisianabelieves.com/docs/default-source/katrina/final-louisana-believes-v5-high-school-performance.pdf?sfvrsn=2.

Maroulis, S., Santillano, R., Jabba, H., & Harris, D. (2015). The Push and Pull of School Performance: Evidence from Student Mobility in New Orleans. *Unpublished manuscript*.

McCaffrey, D.F., T. R. Sass, J. R. Lockwood, K. Mihaly (2009). The Intertemporal Variability of Teacher Effect Estimates. *Education Finance Policy* 4(4):, 572-606.

McLaughlin, K.A., Berglund, P., Gruber, M.J., Kessler, R.C., Sampson, N.A., & Zaslavsky, A.M. (2011). Recovery from PTSD after Hurricane Katrina. *Depression and Anxiety* 28: 439–446.

National Alliance for Public Charter Schools (2013). *A Growing Movement: America's Largest Charter School Communities*. Downloaded July 3, 2015 from: http://www.publiccharters.org/press/students-32-school-districts-attend-public-charter-schools-market-share-report/.

National Commission on Excellence in Education (1983) *A Nation at Risk*. Washington, DC: U.S. Government Printing Office.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, 92(2), 263-283.

Noell, G.H. & Gansle, K.A. (2009). Teach for America Teachers' Contribution to Student Achievement in Louisiana in Grades 4-9: 2004-2005 to 2006-2007. Unpublished manuscript.

Obama, B. (2010). "Remarks by the President on the Fifth Anniversary of Hurricane Katrina in New Orleans, Louisiana." August 29, 2010. Retrieved April 6, 2013 from www.whitehouse.gov.

Pane, J.F., McCaffrey, D.F., Tharp-Taylor, S., & Asmus, G.J., Stokes, B.R. (2006). *Student Displacement in Louisiana After the Hurricanes of 2005 Experiences of Public Schools and Their Students*. Santa Monica, CA: Rand Corporation.

Pane, J.F., McCaffrey, D.F., Kalra, N. & Zhou, A.J. (2008) Effects of Student Displacement in Louisiana During the First Academic Year After the Hurricanes of 2005. *Journal of Education for Students Placed at Risk* 13(2-3), 168-211.

Paxson, C. & Rouse, C.R. (2008). Returning to New Orleans after Hurricane Katrina. *American Economic Review*, 98(2): 38-42.

P.B. v. Pastorek (2010). No. 2:10-cv-04049. The U.S. District Court of the Eastern District of Louisiana. October 25, 2010.

Perry, A., Harris, D., Buerger, C., & Mack, V. (2015). *The Transformation of New Orleans Public Schools: Addressing System-Level Problems Without a System*. New Orleans, LA: The Data Center.

Peterson, P. (2014). Holding Students to Account. In *What Lies Ahead for America's Children and Their Schools*, Chester E. Finn and Richard Sousa (Eds). Stanford, CA: Hoover Institution Press.

Pischke, J-S. (2005). Empirical Methods in Applied Economics: Lecture Notes. Downloaded July 24, 2015 from: http://econ.lse.ac.uk/staff/spischke/ec524/evaluation3.pdf.

Ravitch, D. (2000). The Great School Wars: A History of the New York City Public Schools. Baltimore, MD: Johns Hopkins University Press.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. Journal of Public Economics, 92(5-6), 1394-1415.

Rothstein, J. (2007). Does Competition Among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000). American Economic Review 97(5), 2026-2037.

Rouse, C. (1998). Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics* 113(2): 553-602.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure. *American Economic Journal: Economic Policy*, 5(2): 251-81.

Ruble, W. & Harris, D. (2014). To charter or not to charter: Developing a testable model of charter authorization and renewal decisions. *Journal of School Choice* 8(3), 362-380.

Sacerdote, B. (2012). When the Saints Come Marching In: Effects of Katrina Evacuees on Schools, Student Performance and Crime. American Economic Journal: Applied, 4(1), 109-135.

Sastry, N. & Gregory, J. (2013). The effect of Hurricane Katrina on the prevalence of health impairments and disability among adults in New Orleans: Differences by age, race, and sex. *Social Science & Medicine* 80, 212-129.

Seicshnaydre, S. & Albright, R.C. (2015). *Expanding Choice and Opportunity in the Housing Choice Voucher Program*. New Orleans: The Data Center.

Strunk, K.O. & Grissom, J.A. (2010). Do Strong Unions Shape District Policies? Collective Bargaining, Teacher Contract Restrictiveness, and the Political Power of Teachers' Unions. *Educational Evaluation and Policy Analysis*, 32(3), 389-406.

Tiebout, C., 1956. A pure theory of local expenditures. Journal of Political Economy 64(5), 416-424.

Tyack, D. (1974). The One Best System: A History of American Urban Education. Cambridge, MA: Harvard University Press.

United States Department of Agriculture (2014). *Eligibility Manual for School Meals*.

Vigdor, J. (2008). The Economic Aftermath of Hurricane Katrina. *Journal of Economic Perspectives* 22(4), 135–154.

Walberg, H. (2014). Expanding the options. In *What Lies Ahead for America's Children and Their Schools*, Chester E. Finn and Richard Sousa (Eds). Stanford, CA: Hoover Institution Press.

Weems, C. F., Taylor, L. K., Cannon, M. F.,Marino, R. C., Romano, D.M., Scott, B. G., & Triplett, V. (2010). Post traumatic stress, context, and the lingering effects of the Hurricane Katrina disaster among ethnic minority youth. *Journal of Abnormal Child Psychology*, 38, 49–56.

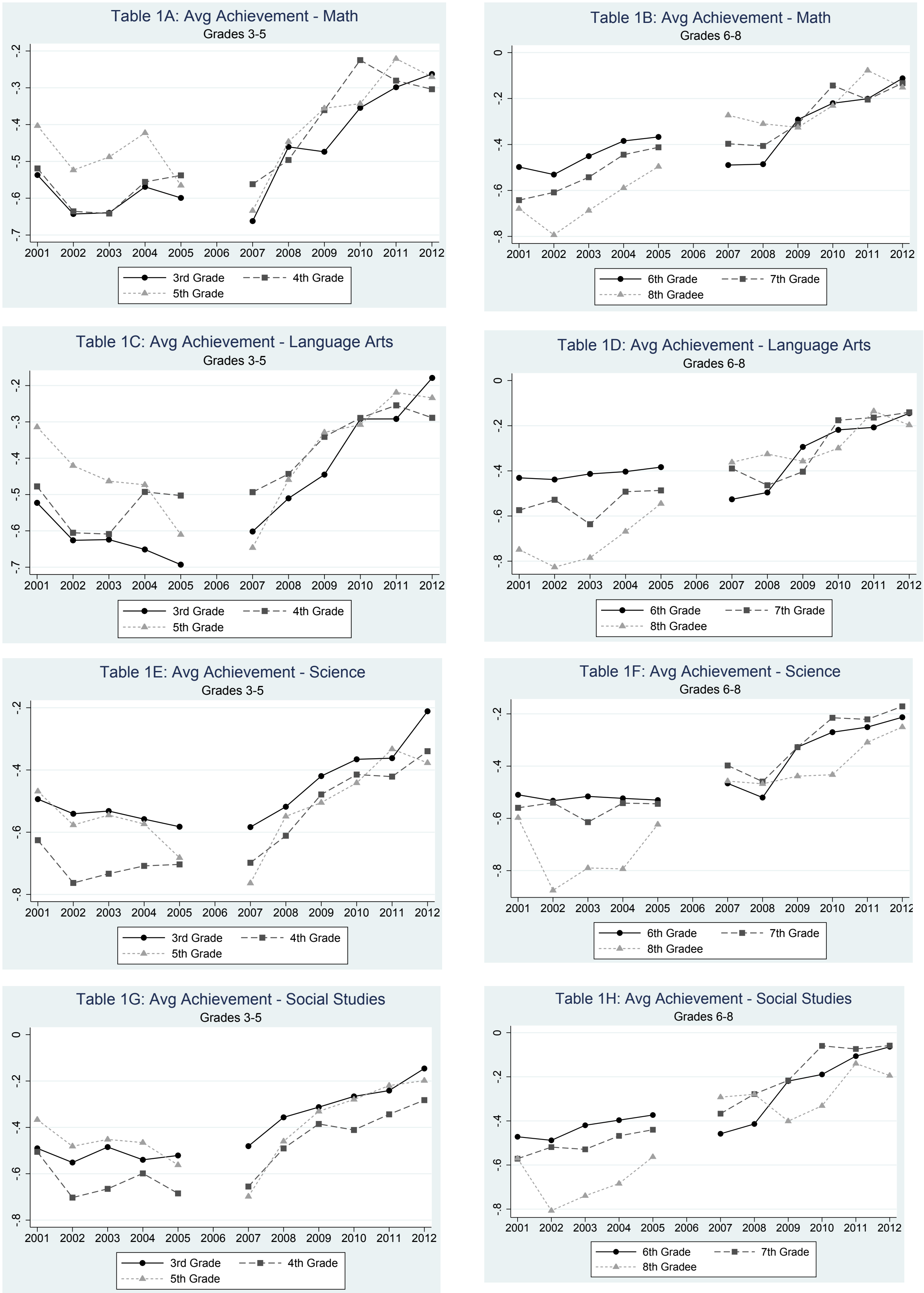
Weixler, L.B., Barrett, N., Jennings, J., Zimmer, R., & Harris, D. (2015). Has the Switch to a Choice System Changed the Distribution of Students by Race, Income, Achievement, and Special Needs Status? Presentation at *The Urban Education Future? Lessons from New Orleans 10 Years After Hurricane Katrina*. June 18-20, New Orleans, LA.

Whitehurst, R. (2012). The Education Choice and Competition Index: Background and Results 2012. Washington, DC: Brookings Institution.

Wong, K. & Shen, F. (2006). Mayors Improving Student Achievement: Evidence from a National Achievement and Governance Database. Paper prepared for the 2006 annual meeting of the Midwestern Political Science Association, April 20-23, 2006.

Xu, Z., Hannaway, J., & Taylor, C. (2007). *Making a Difference? The Effects of Teach for America in High School.* American Institutes for Research, National Center for the Analysis of Longitudinal Data in Education Research.

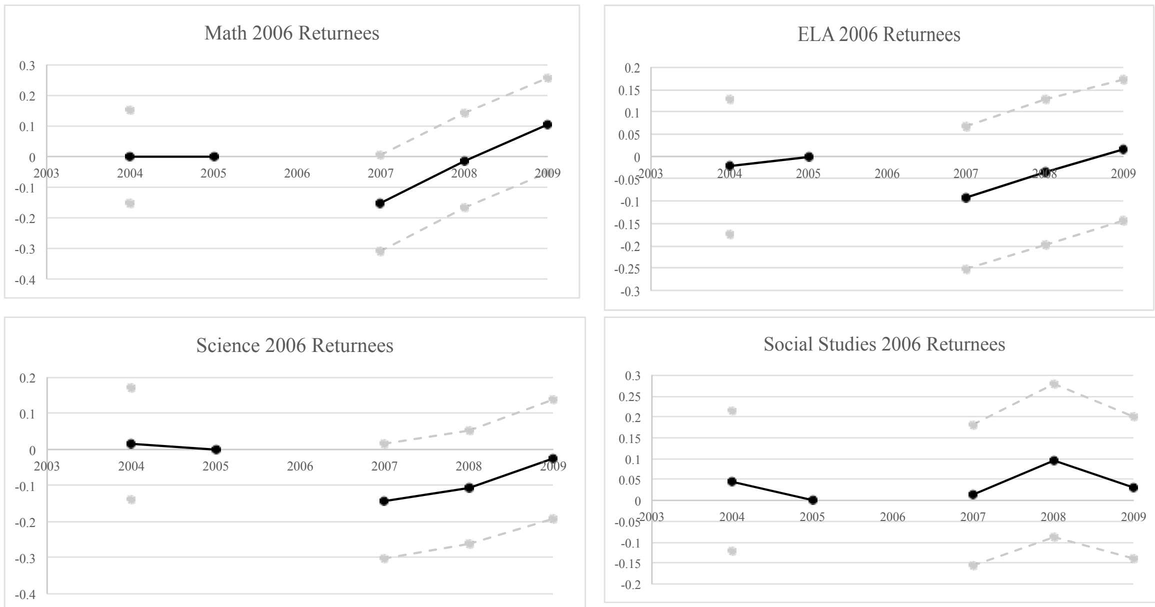
Figure 1: Trends in New Orleans' Student Achievement Levels



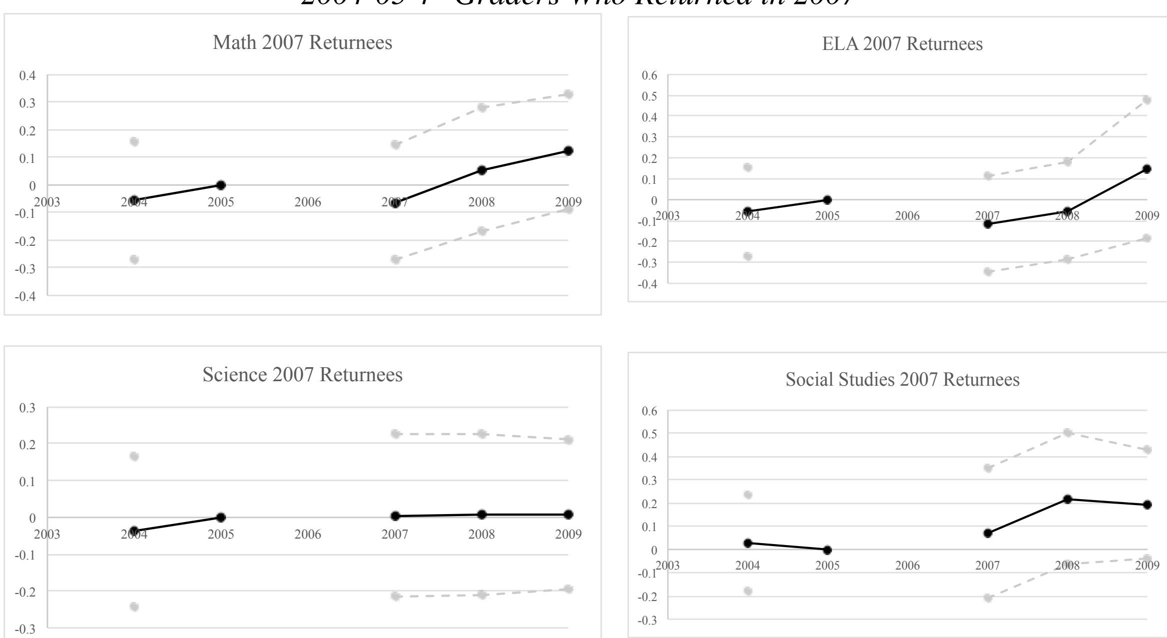
Notes: The y-axis indicates New Orleans test scores standardized so that the statewide N(0,1). The 2005 scores are the last set before the hurricanes and the 2007 scores are the first available in New Orleans post-hurricane.

Figures 2: Panel-M1 Estimates of Average Treatment Effects

2004-05 4th Graders Who Returned in 2006



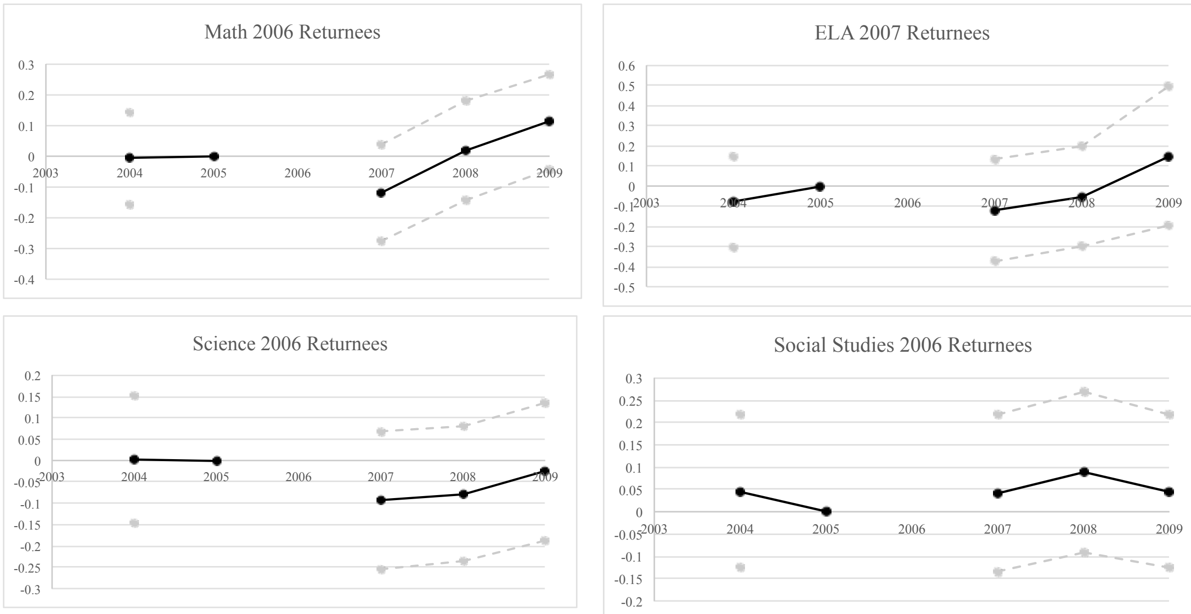
2004-05 4th Graders Who Returned in 2007



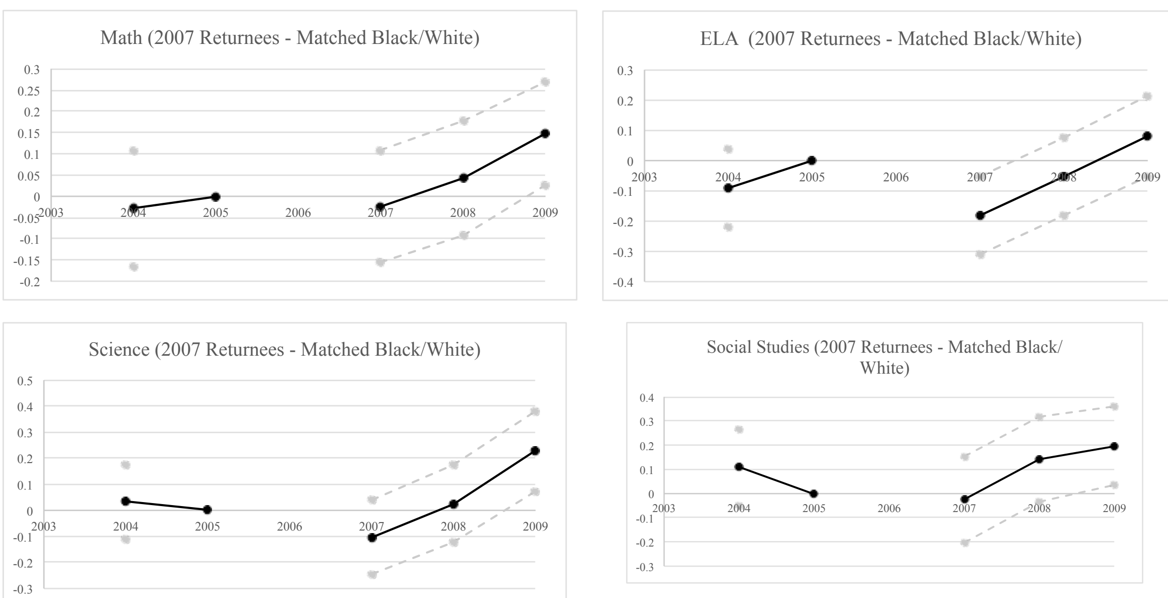
Notes: Results are based on panel estimation of equation (2) using matching method version 1. See additional detail in Table 4.

Figure 3: Panel-M2 Estimates of Average Treatment Effects

2004-05 4th Graders Who Returned in 2006

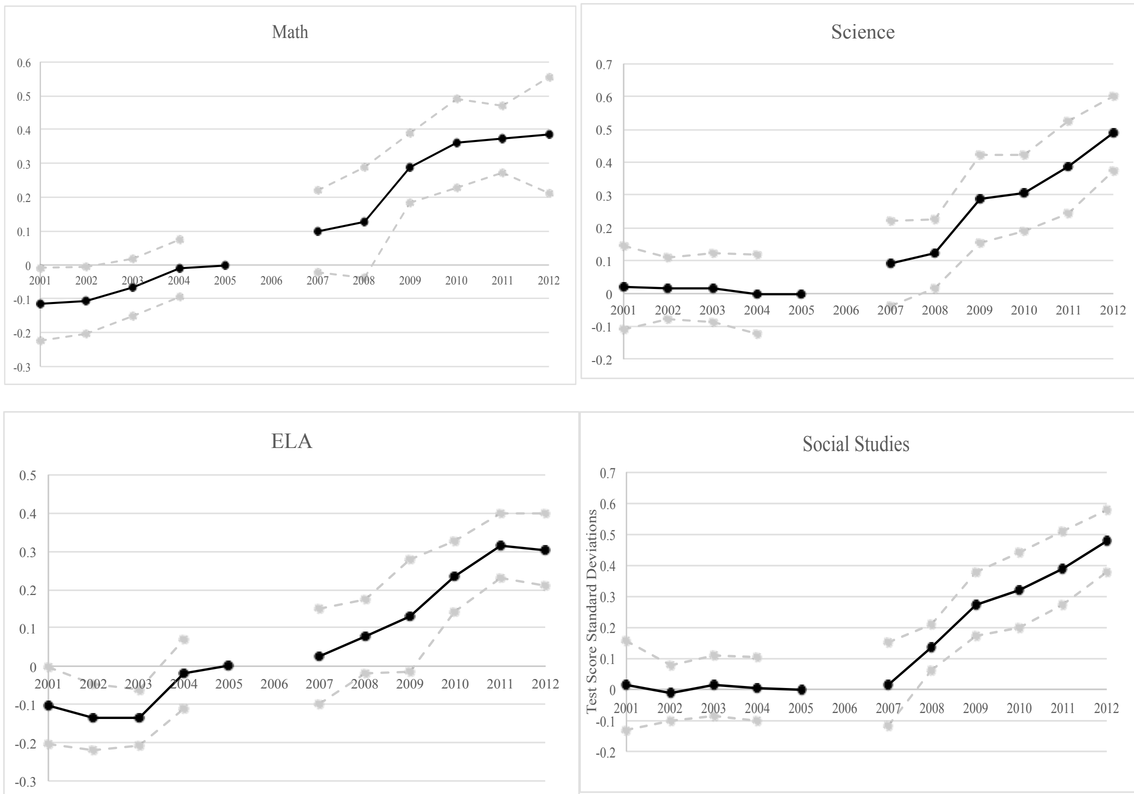


2004-05 4th Graders Who Returned in 2007



Notes: Results are based on panel estimation of equation (2) using matching method version 1. See additional detail in Table 4.

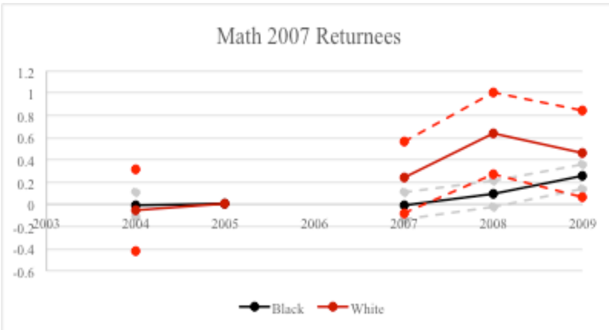
Figure 4: Pooled Estimates of Average Treatment Effects

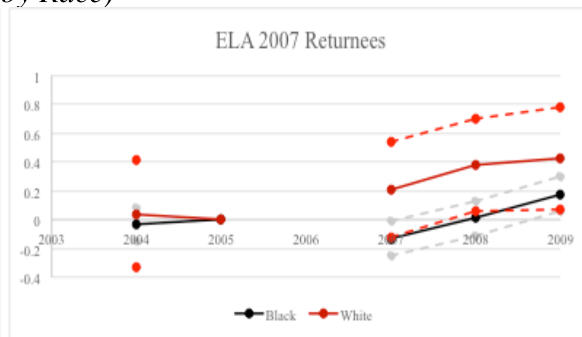


Notes: Effects are averaged across grade levels (weighted). Since these are based on pooled cohorts, and some students are new to the district, they cannot be reported by year of return as they are in Figures 2A-2B. Table 5 model (2) for additional details

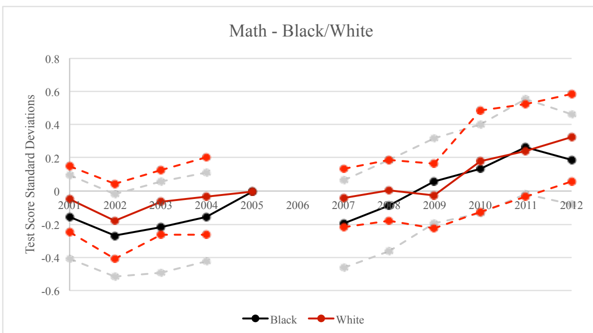
Figures 5: Effect Heterogeneity from Panel Analysis
(4th Grade 2007 Returnees Only)

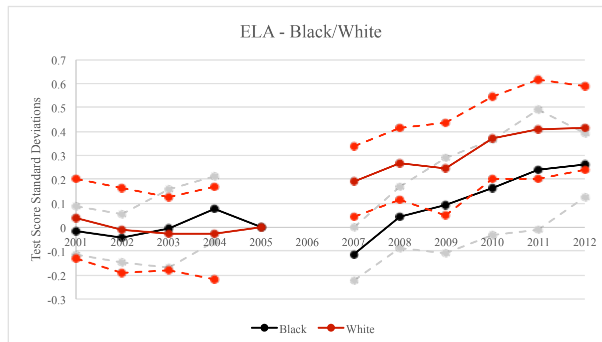
Panel-M2 (by Race)



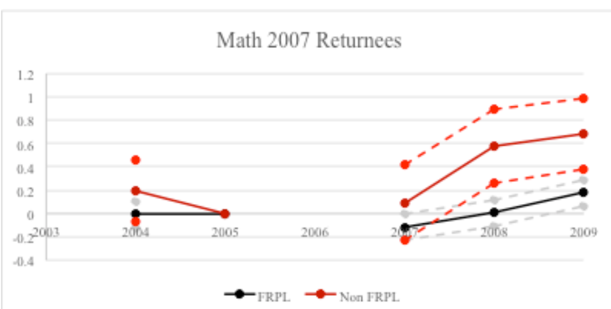


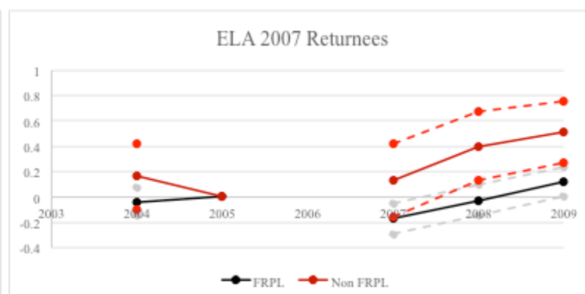
Pooled (by Race)



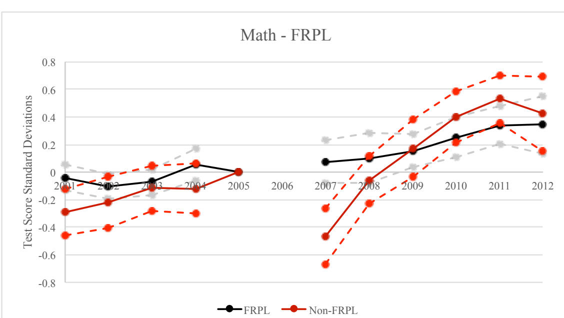


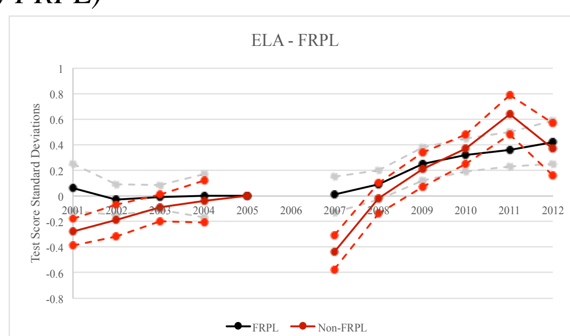
Panel-M2 (by FRPL)





Pooled (by FRPL)





Notes: The panel results are a variation of panel-M1 where the comparison group is stratified on the subgroup rather than matched. [Results for 2006 returnees will be added.]

Table 1: Descriptive Statistics

Panel A: Demographics	2004-05					2011-12					Mean
	N	Mean	s.d.	Min	Max	N	Mean	s.d.	Min	Max	Diff.
African-American	28,063	0.931	0.253	0	1	18,493	0.898	0.302	0	1	-0.033
Hispanic	28,063	0.012	0.111	0	1	18,493	0.026	0.160	0	1	0.014
Other	28,063	0.021	0.144	0	1	18,493	0.024	0.154	0	1	0.003
White	28,063	0.035	0.184	0	1	18,493	0.051	0.220	0	1	0.016
FRL	27,803	0.837	0.369	0	1	18,497	0.867	0.340	0	1	0.030
Special Education	28,073	0.109	0.312	0	1	18,484	0.100	0.300	0	1	-0.009
ELL	28,073	0.019	0.136	0	1	18,500	0.021	0.142	0	1	0.002

Panel B: Test Scores		2004-05					2011-12					Mean
	Grade	N	Mean	s.d.	Min	Max	N	Mean	s.d.	Min	Max	Diff.
Math	3rd	4,281	-0.569	0.988	-3.116	3.118	3,116	-0.263	1.023	-3.496	3.043	0.307
Math	4th	5,841	-0.485	1.097	-4.087	3.249	3,322	-0.3	1.019	-4.226	2.658	0.181
Math	5th	4,483	-0.546	0.935	-2.915	2.868	2,715	-0.27	1.038	-3.354	2.917	0.278
Math	6th	4,403	-0.352	0.952	-2.390	3.032	2,931	-0.11	1.066	-3.344	3.102	0.241
Math	7th	4,162	-0.391	0.998	-2.620	2.910	2,723	-0.133	1.102	-3.419	2.699	0.258
Math	8th	4,729	-0.471	1.172	-4.479	2.892	2,751	-0.160	1.104	-4.958	3.590	0.311
Reading	3rd	4,271	-0.655	0.959	-2.910	2.728	3,118	-0.179	1.072	-3.423	3.339	0.476
Reading	4th	5,843	-0.444	1.095	-3.978	3.313	3,320	-0.289	1.118	-4.223	3.166	0.156
Reading	5th	4,487	-0.588	0.946	-3.060	2.507	2,716	-0.232	1.064	-4.153	3.172	0.356
Reading	6th	4,404	-0.365	0.938	-2.294	2.778	2,931	-0.145	1.031	-3.958	3.888	0.220
Reading	7th	4,161	-0.465	0.980	-2.277	2.712	2,728	-0.140	1.028	-3.933	3.116	0.325
Reading	8th	4,431	-0.522	1.127	-4.466	2.259	2,756	-0.205	1.110	-4.860	3.663	0.317
Science	3rd	4,275	-0.564	0.873	-2.942	3.681	3,106	-0.211	1.014	-4.258	3.858	0.353
Science	4th	5,834	-0.655	1.086	-4.213	3.536	3,319	-0.339	1.020	-4.164	3.083	0.316
Science	5th	4,483	-0.666	0.795	-3.080	2.493	2,713	-0.374	1.084	-4.475	4.222	0.292
Science	6th	4,399	-0.517	0.796	-2.446	2.889	2,935	-0.213	1.014	-4.294	3.962	0.304
Science	7th	4,153	-0.528	0.881	-2.665	2.691	2,722	-0.173	1.042	-4.535	3.878	0.355
Science	8th	4,173	-0.586	1.070	-4.008	2.918	2,723	-0.252	1.058	-4.611	3.902	0.334
Social Studies	3rd	4,278	-0.508	0.979	-3.614	2.874	3,105	-0.146	1.027	-3.885	3.846	0.362
Social Studies	4th	5,827	-0.617	1.220	-4.219	2.895	3,319	-0.282	1.089	-4.571	3.903	0.335
Social Studies	5th	4,483	-0.549	0.917	-3.284	3.138	2,716	-0.196	1.086	-4.283	2.985	0.353
Social Studies	6th	4,400	-0.361	0.905	-2.946	3.504	2,934	-0.062	1.034	-4.087	3.906	0.299
Social Studies	7th	4,155	-0.426	0.894	-3.040	3.099	2,725	-0.061	1.056	-4.310	4.077	0.365
Social Studies	8th	4,147	-0.532	1.120	-3.677	3.790	2,719	-0.195	1.121	-4.319	3.769	0.337

Notes: Table 1 includes New Orleans students in the spring testing file for the given year. The distribution of individual student scores is normalized to N(0,1) for the statewide population within years, grade, and subject. In some cases, the New Orleans standard deviation is above or below that statewide mean. The mean differences indicated changes in the New Orleans population and scores before and after the reforms.

Table 2: Pre-Katrina Mean Differences Between New Orleans and Comparison Groups (2004-05 School Year Characteristics)

Panel A: Demographics		New Orleans		Other Hurricane Districts (Matched)		New Orleans Minus Comparison	
		Panel	Pooled	Panel	Pooled	Panel	Pooled
		(1)	(2)	(3)	(4)	(5)	(6)
African-American		0.929	0.931	0.378	0.688	0.552	0.244
Hispanic		0.010	0.013	0.020	0.020	-0.010	-0.008
Other		0.026	0.021	0.028	0.026	-0.002	-0.005
White		0.035	0.035	0.574	0.266	-0.540	-0.231
FRL		0.878	0.830	0.714	0.799	0.164	0.031
Special Education		0.107	0.109	0.274	0.163	-0.167	-0.054
ELL		0.025	0.019	0.011	0.011	0.014	0.008
Panel B: Test Scores							
Math	3rd	N.A.	-0.569	N.A.	-0.002	N.A.	-0.567
Math	4th	-0.503	-0.485	-0.438	-0.026	-0.065	-0.459
Math	5th	-0.510	-0.546	-0.455	-0.108	-0.056	-0.439
Math	6th	N.A.	-0.352	N.A.	-0.266	N.A.	-0.086
Math	7th	N.A.	-0.391	N.A.	-0.394	N.A.	0.002
Math	8th	N.A.	-0.471	N.A.	-0.305	N.A.	-0.166
Reading	3rd	N.A.	-0.655	N.A.	-0.068	N.A.	-0.587
Reading	4th	-0.481	-0.444	-0.424	-0.044	-0.057	-0.401
Reading	5th	-0.560	-0.588	-0.499	-0.192	-0.061	-0.396
Reading	6th	N.A.	-0.365	N.A.	-0.300	N.A.	-0.065
Reading	7th	N.A.	-0.465	N.A.	-0.321	N.A.	-0.145
Reading	8th	N.A.	-0.522	N.A.	-0.350	N.A.	-0.172
Science	3rd	N.A.	-0.564	N.A.	-0.087	N.A.	-0.478
Science	4th	-0.700	-0.655	-0.599	-0.019	-0.102	-0.636
Science	5th	-0.666	-0.666	-0.611	0.023	-0.055	-0.689
Science	6th	N.A.	-0.517	N.A.	-0.290	N.A.	-0.228
Science	7th	N.A.	-0.528	N.A.	-0.319	N.A.	-0.209
Science	8th	N.A.	-0.586	N.A.	-0.357	N.A.	-0.229
Social Studies	3rd	N.A.	-0.508	N.A.	-0.004	N.A.	-0.504
Social Studies	4th	-0.659	-0.617	-0.561	0.069	-0.099	-0.686
Social Studies	5th	-0.554	-0.549	-0.511	-0.036	-0.043	-0.513
Social Studies	6th	N.A.	-0.361	N.A.	-0.285	N.A.	-0.076
Social Studies	7th	N.A.	-0.426	N.A.	-0.323	N.A.	-0.103
Social Studies	8th	N.A.	-0.532	N.A.	-0.312	N.A.	-0.221

Notes: All data are from the 2004-05 school year. "Panel" specification only includes those students in 4th and 5th grade who eventually return to their 2004-05 school district after the hurricane. "Pooled" specification includes all students in tested grades. In the later panel analysis, we track the 2004-05 4th and 5th graders into the post-Katrina years, which is why test scores for grades 3 and 6-8 are missing in this table. The matched samples in the hurricane-affected districts are weighted by the number of New Orleans students they are compared with. Demographic sample come from those students matched based on math test scores.

Table 3: Effects of Population Change

Panel A: Population Change (Avearge Pre-Katrina 4th Grade Information for Returnees)

	New Orleans			Hurricane-Affected Districts			
	Full Sample	Returnees	Change	Full Sample	Returnees	Change	Diff-in-Diff
FRL	0.886	0.886	0.000	0.615	0.604	-0.011	0.011
Special Ed	0.110	0.098	-0.012	0.167	0.172	0.006	-0.018
ELL	0.018	0.018	0.001	0.026	0.024	-0.002	0.002
Reading Scores	-0.444	-0.478	-0.035	0.171	0.191	0.020	-0.055

Panel B. Census Demographic Changes

	New Orleans			Hurricane-Affected Districts			
	1999	2013	Change	1999	2013	Change	Diif-in-Diff
Income (2013 $)	43,189	42,453	-1.70%	69,659	71,408	2.51%	-4.22%
Prop. BA+	0.10	0.15	0.05	0.16	0.19	0.03	0.02
Prop. Child Poverty	0.57	0.58	0.01	0.30	0.32	0.02	-0.01
Prop. < H.S.	0.33	0.20	-0.13	0.23	0.16	-0.07	-0.06

Panel C. Partial Correlations Between Demographics and Test Scores (from ECLS)

	Dep Var: Test Levels			Dep Var: Test Gains	
	Grade 3	Grade 5	Grade 8	Grade 5	Grade 8
Income (thous., 2013 $)	0.003	0.003	0.003	0.0004	0.0009
	(0.0002)	(0.0002)	(0.0003)	(0.0001)	(0.0002)
BA+	0.139	0.253	0.229	0.046	0.092
	(0.021)	(0.023)	(0.03)	(0.013)	(0.022)
Child Poverty	-0.437	-0.423	-0.402	-0.082	-0.101
	(0.028)	(0.035)	(0.051)	(0.022)	(0.038)
<H.S.	-0.369	-0.366	-0.405	-0.08	-0.076
	(0.044)	(0.048)	(0.065)	(0.029)	(0.054)

Panel D. Predicted Effects of Census Demographic Change on Student Test Scores (Using Panels B and C)

	Test Levels			Test Gains		
	Grade 3	Grade 5	Grade 8	Grade 5	Grade 8	Cumulative
Income (thous., 2013 $)	-0.007	-0.007	-0.007	-0.001	-0.002	-0.012
BA+	0.003	0.005	0.005	0.001	0.002	0.007
Child Poverty	0.004	0.004	0.004	0.001	0.001	0.008
<H.S.	0.022	0.022	0.024	0.005	0.005	0.044
Average	0.005	0.006	0.006	0.001	0.001	0.012

Notes: Panel A shows differences-in-differences (DD) of demographics and test scores (from administrative data) between all pre-Katrina students in the respective districts and the returnees. Panel B shows DD in district-wide demographics based on Census data. Panel C reports regression coefficients based on the federal ECLS, using demographics that line up with the Census measures; we regressed reading score levels (and gains, respectively) on the variable in the left column plus a vector of school fixed effects; each reported coefficient is from a different regression. Standard errors are in parentheses. Panel D provides simulated effects of demographic change; specifically, we inserted the Census-based DD changes from Panel B into the regression model in Panel C. Standard errors of prediction are available upon request.

Table 4: Average Treatment Effects from Panel Analysis, 2006 Returnees

	Whole State	Whole State w/ Student Matching	Hurricane Districts Only	Hurricane Districts w/ Student Matching
	2005 4th Grade Cohort 2005 vs 2009 Diff-in-Diff			
Math				
Post x NOLA	0.210***	0.118*	0.180***	0.100
s.e.	(0.063)	(0.066)	(0.065)	(0.079)
Parallel Trends Test	0.092	0.007	0.169***	0.001
ELA				
Post x NOLA	0.110*	0.044	0.132**	0.013
	(0.064)	(0.067)	(0.066)	(0.081)
	0.240***	0.013	0.207***	0.022
Science				
Post x NOLA	0.196***	0.013	0.189***	-0.027
	(0.064)	(0.067)	(0.066)	(0.084)
	-0.034	-0.018	-0.037	-0.015
Social Studies				
Post x NOLA	0.227***	0.028	0.251***	0.031
	(0.068)	(0.070)	(0.069)	(0.087)
	-0.040	-0.028	-0.065	-0.046
Number of Districts	76	76	8	8
	2005 5th Grade Cohort 2005 vs 2008 Diff-in-Diff			
Math				
Post x NOLA	0.185***	0.075	0.183***	0.035
	(0.068)	(0.071)	(0.070)	(0.085)
	-0.051	-0.000	-0.079	0.008
ELA				
Post x NOLA	0.253***	0.036	0.208***	-0.023
	(0.067)	(0.070)	(0.069)	(0.086)
	-0.243***	-0.008	-0.203***	0.016
Science				
Post x NOLA	0.114*	-0.022	0.108	-0.121
	(0.066)	(0.068)	(0.067)	(0.083)
	-0.020	0.031	-0.063	0.026
Social Studies				
Post x NOLA	0.240***	0.062	0.220***	0.054
	(0.066)	(0.068)	(0.067)	(0.086)
	-0.050	0.011	-0.040	0.020
Number of Districts	76	76	8	8

Notes: The first number in each cell is the point estimate for δ in equation (1) with estimation is at the student level. Each cell represents a separate regression. Standard errors are Huber-White robust. [Estimation based at district-by-year aggregation is forthcoming, although the preferred coefficients are already statistically significant, so will have a minimal influence.] The top portion of each panel pertains to pre-Katrina 4th grade returnees and the bottom portion pertains to pre-Katrina 5th grade returnees (in the respective years). Pre-Katrina 3rd graders are omitted so that parallel trends can be tested. Columns (2) and (4) are weighted by the number of times a student is matched using a Mahalanobis matching process on 2004 and 2005 test scores. See the text for discussion of the matching process. See model (1) for additional details.
*** Signficant at 1%, ** Significant at 5%, * Significant at 10%

Table 5: Average Treatment Effects from Pooled Analysis (2005 to 2012)

Panel A: Math and Reading Avg Test Score Levels				
Math (Post x NOLA)	Whole State	Whole State w/ School Matching	Hurricane Districts	Hurricane Districts w/ School Matching
3rd Grade	0.360***	0.357***	0.310***	0.509**
s.e.	(0.029)	(0.049)	(0.071)	(0.158)
Parallel Trends Test	[-0.028]	[0.012]	[-0.047]	[0.005]
4th Grade	0.243***	0.300***	0.160***	0.362***
	(0.026)	(0.063)	(0.030)	(0.077)
	[0.026]	[0.120]	[0.010]	[-0.050]
5th Grade	0.342***	0.355***	0.256*	0.368
	(0.031)	(0.038)	(0.106)	(0.230)
	[-0.070]	-0.006	[-0.071]	[0.111]
6th Grade	0.299***	[0.289***]	0.265**	0.409
	(0.025)	(0.044)	(0.077)	(0.206)
	[-0.247***]	-0.223***	[-0.206***]	[0.062]
7th Grade	0.335***	0.355***	0.290***	0.267
	(0.022)	(0.033)	(0.049)	(0.197)
	[-0.146***]	[-0.112**]	[-0.161***]	[-0.122]
8th Grade	0.398***	0.524***	0.339***	0.386***
	(0.028)	(0.037)	(0.075)	(0.073)
	[0.053]	[0.170*]	[0.060]	[0.122]
Combined	0.327***	0.366***	0.267***	0.376***
	(0.024)	(0.027)	(0.063)	(0.088)
	[-0.071*]	[0.002]	[-0.073]	[0.018]
Number of Districts	87	56	8	7
ELA (Post x NOLA)				
3rd Grade	0.547***	0.507***	0.515***	0.412***
	(0.028)	(0.035)	(0.059)	(0.103)
	[0.094]***	[0.142***]	[0.091**]	[-0.276**]
4th Grade	0.219***	0.213***	0.180***	0.051
	(0.021)	(0.048)	(0.031)	(0.090)
	[0.046]	[0.108]	[0.085]	[-0.059]
5th Grade	0.428***	0.478***	0.308***	0.360**
	(0.029)	(0.044)	(0.044)	(0.108)
	[0.023]	[0.006]	[-0.020]	[0.009]
6th Grade	0.280***	0.320***	0.223***	0.243
	(0.019)	(0.053)	(0.031)	(0.142)
	-0.212***	[-0.158***]	[-0.216***]	[-0.105]
7th Grade	[0.405***]	0.415***	0.354***	0.414***
	(0.018)	(0.041)	(0.026)	(0.079)
	[-0.076]	[-0.068]	[-0.095]	[-0.007]
8th Grade	0.403***	0.409***	0.358***	0.308***
	(0.018)	(0.031)	(0.052)	(0.074)
	[0.062]	[0.068]	[0.074]	[0.021]
Combined	0.375***	0.380***	0.315***	0.295***
	(0.019)	(0.024)	(0.032)	(0.032)
	[-0.018]	[0.011]	[-0.022]	[-0.052]
Number of Districts	87	56	8	7

Table 5 (continued)

Panel B: Science and Social Studies Avg Test Score Levels

Science (Post x NOLA)	Whole State	Whole State w/ School Matching	Hurricane Districts	Hurricane Districts w/ School Matching
3rd Grade	0.413***	0.384***	0.411***	0.316*
s.e.	(0.023)	(0.050)	(0.059)	(0.166)
Parallel Trends Test	[-0.019]	[0.038]	[-0.032]	[-0.140]
4th Grade	0.402***	0.451***	0.348***	0.582***
	(0.028)	(0.059)	(0.027)	(0.055)
	[0.055]	[0.115*]	[0.083]	[-0.152]
5th Grade	0.363***	0.392***	0.330***	0.550**
	(0.029)	(0.041)	(0.086)	(0.163)
	[-0.022]	[0.023]	[-0.008]	[-0.016]
6th Grade	0.369***	0.331***	0.361***	0.381***
	(0.021)	(0.048)	(0.058)	(0.066)
	[0.007]	[-0.033]	[0.011]	[0.128]
7th Grade	0.441***	0.427***	0.450***	0.485**
	(0.017)	(0.029)	(0.048)	(0.145)
	[0.060]	[0.064]	[0.070]	[0.137]
8th Grade	0.428***	0.448***	0.386***	0.478***
	(0.021)	(0.043)	(0.064)	(0.067)
	[0.128]	[0.112]	[0.130]	[0.052]
Combined	0.406***	0.419***	0.380***	0.483***
	(0.019)	(0.026)	(0.058)	(0.070)
	[0.036]	[0.063*]	[0.043]	[0.012]
Number of Districts	87	56	8	7
Social Studies (Post x NOLA)				
3rd Grade	0.420***	0.396***	0.334***	0.365**
	(0.023)	(0.041)	(0.067)	(0.149)
	[0.050]	[0.043]	[0.048]	[-0.116]
4th Grade	0.415***	0.436***	0.376***	0.554***
	(0.028)	(0.064)	(0.026)	(0.093)
	[0.109]	[0.169**]	[0.120]	[-0.022]
5th Grade	0.421***	0.424***	0.369**	0.579***
	(0.034)	(0.047)	(0.109)	(0.085)
	[-0.081]	[-0.003]	[-0.097]	[-0.006]
6th Grade	0.357***	0.234***	0.348***	0.435***
	(0.025)	(0.032)	(0.072)	(0.013)
	[-0.179***]	[-0.154***]	[-0.187***]	[-0.006]
7th Grade	0.441***	0.400***	0.433***	0.438**
	(0.022)	(0.051)	(0.072)	(0.120)
	[0.016]	[-0.031]	[-0.014]	[-0.142]
8th Grade	0.428***	0.413***	0.391***	0.429***
	(0.022)	(0.040)	(0.050)	(0.037)
	[0.203]	[0.128]	[0.247*]	[0.086]
Combined	0.417***	0.380***	0.388***	0.471***
	(0.022)	(0.030)	(0.060)	(0.054)
	[0.022]	[0.044]	[0.021]	[-0.032]
Number of Districts	87	56	8	7

Notes: Coefficients are the estimation of δ in equation (1) with test score level as the dependent variable and aggregation to the district-by-year level. Huber-White standard errors in parentheses. Only 2004-05 and 2011-12 scores are included. Columns (1) and (3) are weighted by district size. Columns (2) and (4) are weighted by district size where the weights come from Mahalanobis matching at the school level of average test scores in 2002. The third row [in brackets] is the coefficient from the parallel trends test where asterisks indicate rejection of the null. See Figures 3A-3D for additional evidence on pre-trends. Significance levels: *** = 0.001, ** = 0.01, * = 0.05.

Table 6A: Annualized Average Treatment Effects based on Students Switching Districts (Switcher-M1)

	Switch In	Switch Out
Math		
Post-Katrina	0.070***	-0.034
	(0.024)	(0.025)
	{0.037}	{0.046}
ELA		
Post-Katrina	0.093***	-0.044*
	(0.024)	(0.025)
	{0.017}	{0.025}
Science		
Post-Katrina	0.086***	0.010
	(0.025)	(0.026)
	{0.043}	{0.039}
Social Studies		
Post-Katrina	0.107***	0.052*
	(0.028)	(0.029)
	{0.028}	{0.039}

Notes: We regress achievement on lagged achievement, grade fixed effects, and an indicator for whether the switch occurred before or after Katrina (*Post-Katrina*), at the student level (no aggregation). Pre-Katrina district switches are included for 2003-2005 and the post-Katrina years are 2010-2012. Robust standard errors are provided in parentheses. In brackets underneath are standard errors clustered at the sending district level for in-switches and the receeiving district levels for out-switcher. The number of observed switches ranges from 3,985 to 4,742. See text and earlier footnotes for more details on the model.

Table 6B: Annualized Average Treatment Effects based on Students Switching Districts (Switcher-M2)

	Switch In	Switch Out
Math		
Post-Katrina	-0.081***	-0.082***
	(0.006)	(0.006)
	{0.013}	{0.014}
Switch Type	-0.065***	-0.122***
	(0.017)	(0.013)
	{0.017}	{0.024}
Switch Type*Post-Katrina	0.143***	0.042
	(0.025)	(0.026)
	{0.041}	{0.024}
ELA		
Post-Katrina	-0.069***	-0.067***
	(0.006)	(0.006)
	{0.010}	{0.014}
Switch Type	-0.095***	-0.107***
	(0.017)	(0.013)
	{0.020}	{0.024}
Switch Type*Post-Katrina	0.151***	0.026
	(0.024)	(0.025)
	{0.025}	{0.025}
Science		
Post-Katrina	-0.067***	-0.073***
	(0.006)	(0.006)
	{0.017}	{0.015}
Switch Type	-0.176***	-0.179***
	(0.018)	(0.014)
	{0.023}	{0.026}
Switch Type*Post-Katrina	0.145***	0.074***
	(0.025)	(0.027)
	{0.049}	{0.040}
Social Studies		
Post-Katrina	-0.058***	-0.067***
	(0.007)	(0.007)
	{0.019}	{0.016}
Switch Type	-0.143***	-0.212***
	(0.020)	(0.014)
	{0.031}	{0.023}
Switch Type*Post-Katrina	0.155***	0.111***
	(0.028)	(0.029)
	{0.036}	{0.046}

Notes: We regress achievement on the variables shown, plus lagged achievement and grade fixed effects. Our estimate of the reform effect comes from the interaction term. Robust standard errors are provided in parentheses. In brackets underneath are standard errors clustered at the sending district level for in-switches and the receeiving district levels for out-switcher. The number of observations is much larger here (60,891-61,754) than in Table 6A because all district switches are included, regardless of whether they involved New Orleans. See text and earlier footnotes for more details on the model.

Table 7: Effect Summary, Bounds, and Cost-Benefit Analysis

Effect Category	2007	2008	2009	2012
Total NOLA improvement rel. to state	0.10	0.15	0.25	0.40
Threats to Identification				
Population Change [1]				
Pre-Kat Scores of Returnees	0.10	0.06	0.04	-0.06
Census/USDOE Simulation	NA	NA	NA	0.01
Interim Schools/Trauma (Pane et al. 2008)	-0.06	NA	NA	NA
Effects from Panel DD				
Panel-M1	-0.07	-0.05	0.08	[0.29]
Panel-M2	-0.05	-0.04	0.04	[0.18]
Effects from Pooled DD				
Table 5	0.08	0.11	0.30	0.40
Lower Bound - 1				0.20
Lower Bound - 2				0.23
Upper Bound				0.40
Dosage (Post-Reform Years in NOLA)				4.50
Annual Cost/Pupil				$1,000
Adjusted Effectiveness/Cost Ratio (ECR)				
Lower Bound - 1				1.45
Lower Bound - 2				1.71
Upper Bound				2.97
Break-Even ECR (Harris, 2009)				0.26
ECR: Preschool				0.30
ECR: Class Size (STAR)				1.58

Notes: "Total Improvement" is based on the trends in Figures 1A-1H. Values for "Pre-Kat Score of Returnees" for 2007-2009 are similar to Table 3 Panel A except that are simple differences for New Orleans (no comparison group). Lower Bound-1 is based on the pooled analysis with an adjustment for estimated bias; it assumes: (a) average of Panel-M1 and Panel-M2 is unbiased; and (b) pooled bias is fixed in magnitude over time. Lower Bound-2 is based on linear projection of early panel results into the future; it assumes: (a) average of Panel-M1 and Panel-M2 is unbiased; and (b) effects continued on the same path after 2009 (consistent with every panel and pooled analysis). Upper Bound is based strictly on the pooled estimates and assumes they are unbiased. All estimates assume no long-term effects of trauma and disruption. Values in brackets are based on projections; all others are actual. "Break-Even ECR" is the effectiveness-cost ratio for the reforms, assuming that a one standard deviation increase in test scores increases future earnings by eight percent and a three percent discount rate. NA = Not Available