Technical Report

THE EFFECTS OF THE NEW ORLEANS POST-KATRINA SCHOOL REFORMS ON STUDENT ACADEMIC OUTCOMES

EDUCATION RESEARCH ALLIANCE FOR NEW ORLEANS

Douglas N. Harris, Tulane University
Matthew Larsen, Lafayette College

February 10, 2016

EducationResearchAllianceNOLA.com

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

Douglas N. Harris
Matthew F. Larsen

February 10, 2016

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, distortions in test scores from high-stakes accountability, influence of the interim schools attended by evacuated students, and the trauma and disruption from the hurricane itself. With the possible exception of test score distortions, 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 on achievement 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, Lindsay Bell Weixler and seminar participants at the National Bureau of Economic Research, Syracuse University, Texas A&M 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.

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 one 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; Abdulkadiroğlu et al., 2014), private school vouchers (Rouse, 1998; Krueger & Zhu, 2004; Abdulkadiroğlu et al., 2015), 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. With autonomy from district and union rules, they would also have the opportunity to meet accountability demands, yielding greater efficiency.

Though the word accountability has been commonly used, the actual incentives and autonomy have changed less 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 local school board politics, unions, and school attendance zones still heavily influenced school operations (Ravitch, 2000). NCLB increased the volume of testing, 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. At about the time Katrina made landfall, only two percent of U.S. students attended charter schools and 13 percent of

[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).

U.S. students attended a non-assigned publicly funded school (Harris & Witte, 2011).[4] 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 consistently competitive effects on nearby 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 much further with school accountability and autonomy than perhaps any district or state ever had. 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 0.35 standard

[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). The percentage of students in charter school has since increased to almost five percent (NCES, 2015).
[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 New Orleans RSD and OPSB schools.

deviations from 2004-05 to 2011-12 (see Table 1). These positive trends, combined with evidence that charter schools in New Orleans (Abdulkadiroğlu et al., 2014) 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 (Dreilinger, 2014) 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 trends and positive charter effects provide 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) 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 out several threats

[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 does not address their post-Katrina improvement.

to identification, including population change, trauma and disruption from the hurricane, the effectiveness of the interim schools that evacuated students temporarily attended, and strategic 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 examines 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, there are five key threats that serve as alternative potential causes of the positive test score trends.

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

Second, when Louisiana families evacuated, they generally placed their children in schools near their temporary residences. 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, then some of the later

increases in achievement might reflect the performance of these interim schools rather than the New Orleans reforms.

Third, prior research has shown that accountability induces some schools to manipulate high-stakes measures and/or reallocate resources in ways that reduce unobserved outcomes that are lower-stakes (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 where accountability pressures are especially high.

Fourth, NCLB had been adopted a few years prior to Katrina and the law's key provisions were about to be implemented. Since low-performing schools are the focus of NCLB sanctions, the post-Katrina improvements in New Orleans' 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 fifth threat could have the opposite influence: Hurricane Katrina was one of the worst disasters in American history[8] and created persistent trauma and anxiety for residents (DeSalvo et al., 2007; Elliott & Pais, 2006; 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 years after the hurricanes

[8] 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).

(Brown et al., 2011)[9] and this apparently reduced academic learning at least in the short term (Pane et al., 2006, 2008; Sacerdote, 2012).

Estimation Strategy

We use a combination of matching and difference-in-difference (DD) analysis to address all of 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):

$$A_{ijt} = \gamma_j + X_{ijt}\beta + \lambda d_t + \delta\left(NOLA_j \cdot d_t\right) + \varepsilon_{ijt} \quad (1)$$

where A_{ijt} is the achievement of student *i* in school district *j* at time *t*, γ_j is a vector of school district fixed effects, X_{ijt} is a vector of student covariates[10], d_t indicates whether the outcomes pertain to a single pre-treatment period or a single post-treatment period, and $NOLA_j$ is an indicator set to unity for New Orleans and zero for students in the matched comparison group districts. No other district in Louisiana experienced the reforms and these therefore serve as a useful comparison group. Under certain assumptions discussed below, 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.[11]

Our first estimates are based on equation (1) using only 2005, the year prior to the reforms, and 2009 or 2012, the most recent post-reform period available in the data

[9] 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).

[10] These include race, free/reduced price lunch status, special education status, limited English proficiency, and grade repetition. In addition, we include bin indicators for each stratum in the matching process.

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

depending on the analysis.[12] There are many reasons to expect, however, that the effects of the reforms emerged gradually over time. In creating an entirely new system of schooling, New Orleans leaders not only had to create new schools, but an entirely new governance structure and new institutions to recruit and develop charter school operators (e.g., 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 traditional school district to reach equilibrium. Given all the changes that occurred, this appears to be a realistic assessment.

To estimate these dynamic effects and 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:

$$A_{ijt} = \gamma_j + \lambda_t + X_{ijt}\beta + \sum_{\tau=0}^{m} \delta_{-\tau}\left(NOLA_j \cdot d_{j,-\tau}\right) + \sum_{\tau=1}^{q} \delta_{+\tau}\left(NOLA_j \cdot d_{j,+\tau}\right) + \varepsilon_{ijt} \quad (2)$$

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

[12] We refer to the spring of the school year throughout the remainder of the study, since this is when students take the tests. So, 2005 means the 2004-05 school year and so on.

[13] When there are more than two periods of data, it is sometimes recommended to add group-specific time trends as follows: $A_{ijt} = \gamma_{0j} + \gamma_{1j}t + \lambda_t + X_{ijt}\beta + \delta(NOLA \cdot d_t) + \varepsilon_{ijt}$where t is a continuous time period variable and γ_{1j} is the slope (Angrist & Pischke, 2009). This specification yields biased estimates, however, when there are dynamic effects (Pischke, 2005). Equation (2) avoids this problem.

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).

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 2005 cohort of 4th graders with the 2012 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 subsample 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 that are of primary interest. With 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

[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.

identification discussed above (population change, strategic behavior, effects of other policies, interim school effects, and trauma/disruption), we take additional steps as well.

Our preferred results are estimated 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) and our preferred estimates include only eight. Aggregation of data to the district-by-year level is an alternative and generally yields conservative standard errors (Bertrand et al., 2004; Angrist & Pischke, 2009). In a few cases, GEE clustered standard errors are larger than Huber-White standard errors and in those cases we report the latter to be conservative. We also estimate the models using estimation with district-level aggregation, which yield even more conservative standard errors (available upon request).[15] The results are robust to these and many other robustness checks.

Data and Matching

The Louisiana Department of Education (LDOE) provided student-level longitudinally linked data for essentially all public school students in the state for each year 2002-2012. Key variables include student test scores, demographics, grade level, and the schools where students enrolled. Pre- and post-Katrina, students took state

[15] A third common alternative, the wild bootstrap (Cameron, Gelbach, & Miller, 2008) is infeasible in this case because there is only one treatment cluster.

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 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, which 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.

The year 2012 is a convenient end point because most of the major reforms were completed by this point and the system had re-stabilized in the number of schools and students.[17] We are examining post-2012 effects and non-test outcomes in ongoing research.

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 just hurricane-affected districts. If the trauma/disruption and interim school effects were the same in New Orleans and other hurricane-affected districts, then this sample

[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.

[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.

restriction 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.[18] Therefore, we view the comparison of the statewide and hurricane-affected districts as only a test for whether trauma/disruption played a role.

In the panel analysis, our first preferred matching method involves the following steps: (a) restrict to hurricane-affected school districts (see above); (b) from those affected districts, 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.[19] Step (b) helps ensure that the comparison group is similar to New Orleans in the unobserved factors associated with return to the original district.[20]

---

[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. They 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.
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.

[19] 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.)

[20] 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 family characteristic 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. People 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. Matching based on year of return helps account for these potentially important differences.

An alternative matching method is identical to that above but also stratifies on one demographic measure (free/reduced price lunch status). This is based on prior evidence that achievement growth varies by student background.[21] Since the results are robust to this alternative matching method, we report results based on test-only matching.

Pooled Matching. For the pooled cross sections, the matching process differs because some of the cohorts are from the post-Katrina period and we can only match on pre-reform outcomes. 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 among post-Katrina cohorts in those schools after the hurricane. With this assumption, we can still match the post-reform cohorts but without relying on post-reform data.

Specifically, for our preferred pooled analysis, we match the post-reform cohorts on pre-reform measures as follows: (a) 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 remaining schools, use Mahalanobis matching to identify comparison schools with composite test score levels in 2002.[22] Note that step (b) applies only to the comparison group; that is, all post-Katrina New Orleans schools count toward New Orleans[23] outcomes regardless of whether they existed pre-Katrina.[24]

[21] Later, in Table 3, we provide direct evidence that demographics are associated with achievement levels and growth. 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.

[22] We match on 2002 only here because this yielded a more valid comparison group compared with other methods (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-treatment patterns and effect estimates were unaffected.

[23] When we say "New Orleans schools," we mean all publicly funded and governed schools in the city, including those authorized by both the RSD and OPSB.

The differences in matching also highlight a potential 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.

With both panel and pooled matching methods, we weight comparison group students/schools based on the number of times they are matched to New Orleans students, so that the weighted distribution in each comparison district is as similar as possible to New Orleans. In the panel analysis, this implies that the weighted number of students is the same in every district cluster 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 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.[25]

---

[24] 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.

[25] 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 treatment unit; and (b) there are smaller units (schools) nested within the treatment unit and most of the variance in outcomes is between these smaller units. 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

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). Below, 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 reduced price lunch (FRPL); almost all the students are racial/ethnic minorities and 93 percent are black. The differences between 2005 and 2012 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 our preferred matching process for the panel analysis. The matching succeeded in finding matched samples of students in hurricane-affected districts that, prior to Katrina, had test score levels similar to New Orleans. Column (5) shows that the panel comparison group is 0.15 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,

district level after doing Mahalanobis matching at the school level, but the Mahalanobis matching removes so much of the pre-treatment variation between districts that this additional step does not improve the match very much.

the pooled matching method yields a difference between New Orleans and the comparison group of 0.34 standard deviations. This of course focuses on test levels, though we are most concerned with the parallel trends tests shown 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.

Panel A of Table 3 provides data on pre-Katrina 3rd graders, including all pre-Katrina students and just those who returned, for New Orleans and other hurricane-affected districts. By 2010, New Orleans returnees had somewhat lower pre-Katrina scores than the overall pre-Katrina New Orleans population, while in the other districts, the returnee scores were higher than the pre-Katrina population. The DD therefore favors the comparison districts by 0.043 standard deviations. In other words, the change in the population reduced New Orleans scores by a small amount.

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 detailed demographics for households with students in

public schools, for each district in the state.[26] Panel B of Table 3 shows that some Census 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).[27] However, the percentage of the population with a BA or higher increased by five percentage points in New Orleans but by only 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 and each of the demographic measures.[28] 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.[29] The results are shown in Panel D. The simulated cumulative effect across 4.2 years in the reformed school system (our estimate of the "dosage"), averaged across the demographic measures, is 0.012 standard deviations; the largest estimate is 0.044 standard deviations.[30] This implies a possible upward bias in pooled analysis, but an extremely small one.

---

[26] 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 publicly available data for the entire school-age population (public and private schools) using all districts in the state as well as other hurricane-affected districts.

[27] The absolute decline in socio-economic characteristics in New Orleans is corroborated by Vigdor (2008).

[28] In each regression, the ECLS test score (in levels and growth, respectively) is regressed on one demographic measure and a vector of school fixed effects.

[29] 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.

[30] 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.

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. For example, the black middle class, whose children also attended public schools in large numbers, also saw a large drop (Plyer, Shrinath, & Mack, 2015). 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.[31] This evidence suggests that population change is not a major threat to identification in the pooled analysis.

Results

Panel Estimates of Average Treatment Effects

Tables 4A and 4B report results from the panel analysis estimation of average treatment effects (ATEs) based on equation (1) for 4th and 5th graders by year of return, separately for 2006 returnees (Table 4A) and 2007 returnees (Table 4B). These tables use the first matching method, which matches on test scores only, but then controls for student demographics, grade repetition, and bin indicators in the effect estimation.[32]

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

[32] The year of return is based on the year the tests were taken, so a 2007 returnee likely returned in the fall of 2006. However, the 2006 returnees almost all returned in spring of 2006 because all the schools were closed through fall of 2005. 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.

The column (1) sample includes almost all Louisiana students who have data pre- and post-hurricane (without matching)[33]; column (2) includes the entire state matched on test 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. Since our test scores end in grade 8, we can follow pre-Katrina 4th (5th) 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).

Half of the 64 estimates are positive and significant and all but seven have positive point estimates.[34] The estimates are similar between the state and hurricane-affected districts, though matching reduces the coefficient magnitude and the increases the likelihood of passing the parallel trend tests (p-values shown under standard errors).

Focusing just on our preferred specification where students are matched to those in other hurricane-affected districts, the point estimates average about 0.12 standard deviations through 2009 for pre-Katrina 4th graders (top of Table 4A). Since the matching is based on test score levels, the sensitivity to matching may mean that NCLB, other statewide policies or a change in the test scale had particular influence on low-performing students and schools in other parts of the state that form the matched sample. In all cases with the hurricane-affected matched comparison, we pass a parallel trends test for the two years prior to the hurricane.

[33] 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.

[34] The percent statistically significant is slightly lower when estimated with the data aggregated to the district-by-year level.

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, by construction, the same as the top of Table 4A for 2009. There are signs, especially in math and ELA, 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.[35] It is difficult to empirically distinguish between these alternative theories, though the results that follow do suggest that schools steadily improved after 2009.[36] The results are generally similar when we switch to matching on demographics.[37] Since there is no clear preferred matching method, we establish bounds later by using the average of the various methods.

Overall, the vast majority of coefficients in Tables 4A and 4B are positive (one-third of those are precisely estimated), and the estimated effects are consistently larger for students who have more post-Katrina years to experience reform effects. Also, in all but one of the 16 cohort-by-subject analyses in Figures 2 and 3, we see a positive trajectory over time in the point estimates.

[35] 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).

[36] The results for the 2005 5th graders are available upon request. They display the same general upward pattern, though it is flatter. 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. The effects for pre-Katrina 5th graders are smaller and include the only two cases in this study where we find negative and significant coefficients, which is partly why we downplay them here. The less positive effects for pre-Katrina 5th graders may be also due to observing fewer years under the reforms (lower dosage).

[37] We also carried out placebo tests and these yielded similar results (available upon request).

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 and 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 and less for later returnees). Finally, the initial dip in scores upon return suggests that trauma/disruption effects may have been larger for New Orleans students and pulled down the measured effects in the short term. If the objective of estimating the long-term cumulative effects of the program, then these panel analysis limitations imply that the estimates in Figure 2 and Table 4 are biased downwards. The analysis that follows avoids these limitations, though may suffer from others.

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. With this pooled method, we can look at longer-term estimates through 2012, three years later than the panel analysis. Again, these are cumulative effects and students enrolled in New Orleans taking the test in 2012 averaged 4.2 years under the reformed system.

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 0.30-0.47 standard deviations across subjects.[38] As in the panel analysis, Figure 4 also suggests that the positive effects are the result of steady improvement leading up to 2012. The estimates also pass a parallel trends test in three-quarters of the grade-by-subject estimates.[39]

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 population change were the driving force behind the estimated 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 have been too young to spend much time in interim schools in the hurricane aftermath. 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.

[38] 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.

[39] 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.

We report results separately by grade level as a test for whether the effects were larger in 4th and 8th grade where the stakes are higher for students. Also, in the both the panel and the pooled, we report results separately by subject since the stakes for schools are somewhat higher for math and reading. We see no evidence that the effects are systematically larger when the stakes are higher, suggesting that strategic behavior and test distortions do not the primary driver behind these estimated effects.[40]

Robustness Checks and Additional Identification Strategies

Identification from District Switchers. A third identification strategy involves 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. In addition to this Switcher Method 1 (M1), we also estimate a Switcher-M2 that accounts for changes in statewide trends in cross-district mobility.[41]

---

[40] We also considered whether the effects got gradually larger across grades since older students have somewhat higher reform dosage than younger students. However, the differences in dosage are slight; 3rd graders had an average dosage of 3.7 years while 8th graders averaged 4.4 years. For 3rd graders, the dosage calculation includes mostly non-tested grades (K-2) since that most of the reform policies, with the exception of test-based accountability, applied to all grades. The similarity in dosage likely explains why the ATEs do not display a clear pattern across grades.

[41] 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 Switcher-M1 model 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). We are interested in $\beta_1$which simply compares 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. We thank Andrew McEachin for suggesting this general approach.

One advantage of the switcher method, like the pooled analysis, is that it allows us to test for effects many years later; we specifically use data from 2003-2005 and 2010-2012. The identifying assumption of the simpler switcher 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 of the same magnitude as the out-switcher effect, but with the opposite sign.

The treatment effects from the switcher methods are in terms of annual growth. To compare these with our own results, which to this point have been cumulative across years, we re-estimated our prior models (equation (1)) with achievement *gains* as the dependent variable. If the switcher analysis is well identified, the differences in magnitudes between the in-switcher and out-switcher coefficients should be similar to the effect estimates from the pooled analysis. This is what we find, i.e., annualized effects of 0.05-0.10 standard deviations. There are two reasons to downplay the switcher results. First, the identifying assumption does 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 just 10 percent of New Orleans students, a small and possibly unusual sample.

Identification from Instrumental Variables. We also carried out an instrumental variables (IV) strategy akin to Abdulkadiroğlu et al. (2014). In short, the second stage regresses achievement growth (pre- versus post-treatment) on the number of years

students attended New Orleans public school post-Katrina.[42] Since post-Katrina attendance is clearly endogenous, we can use pre-Katrina New Orleans public school enrollment as an instrument.

One important limitation of the IV method is that the exclusion restriction is implausible since attendance in New Orleans public schools pre-Katrina could directly influence post-Katrina outcomes, though the weaker form of the assumption required in the Abdulkadiroğlu et al. (2014) method lessens the problem somewhat. In this empirical context, unlike theirs, an additional issue arises: In our panel analysis we stratify by year of return to reduce trauma and interim school effects (see footnote 20). This is not possible under their IV method, which indirectly incorporates the year of return through the continuous dosage variable.[43] In any event, the IV estimates effects are similar but somewhat less positive than our panel estimates and usually statistically insignificant.

Other Robustness Checks. We generally estimated and reported effects based on student-level data disaggregation with GEE standard errors clustered at the district level. While this approach has the advantage of allowing us to include student-level covariates and bin indicators, the small number of clusters calls into question the GEE assumptions.[44] We therefore re-estimated the models by aggregating up to the district

[42] Specifically, our second stage equation is: $Y_{ij}^{2009} - Y_{ij}^{2005} = \gamma_j + X_{ijt}\beta + \delta D_{ijt} + \varepsilon_{ijt}$ where D_{ijt} is the number of years spent in a New Orleans public school under the reforms. The first stage is: $D_{ijt} = X_{ijt}\beta + \pi Z_i + \nu_{ijt}$ where Z_i is the instrument that indicates whether student i attended a public school in New Orleans in 2005. The estimates easily satisfy the first stage. (As above, bin indicators are considered part of X_{ijt} and are not shown explicitly in the equations.) We thank Joshua Angrist for noting the similarity between our situation and theirs and for suggesting this approach.

[43] This is not an issue in the Abdulkadiroğlu et al. (2014) empirical context where there are no plausible trauma and interim school effects. We considered an alternative version of their IV method where we stratify on year of return, but rely only on post-return variance in the dosage, but there is almost no such variance to speak of during this short panel.

[44] 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, omitting the student-level covariates and bin indicators. This had only a minimal influence on either the point estimates or the standard errors (available upon request). As noted earlier, the effects are also qualitatively similar when switching the dependent variable to achievement growth (a form of triple difference).[45] Finally, we find no evidence of bias from missing data.[46]

ATE Bounds

Our objective is to estimate the average treatment effects (ATEs) of the New Orleans school reforms through 2012. Given the varied methods and years reported above, we calculate bounds using extreme sets of assumptions about the previously reported results. Our first lower bound estimate is based on the pooled analysis but using the panel results to estimate bias. It assumes (a) that the difference between the pooled estimates and the panel estimates identifies bias in the pooled estimates[47]; and (b) that the pooled bias in 2009 remains of the same magnitude afterwards. The average bias using this method is 0.11 standard deviations. Subtracting this from the pooled estimate of 0.40

---

level demographics over time, which are extremely small and have little variance across districts, resulting in implausible parameter estimates.

[45] 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 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.

[46] 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.

[47] Though not shown elsewhere, we show in Table 7 the results using both our main matching method from Tables 4A and 4B and the alternative method where we match not only on test scores but student demographics. This highlights the similarity in results between the two methods and provides additional basis for establishing bounds.

standard deviations yields an ATE of 0.29.[48] With estimation at the district level of aggregation yielded an ATE of 0.20 standard deviations, which is what we report in the conclusion.

The second lower bound is based on linear projection of the panel results to 2012; it assumes that: (a) the panel results are unbiased estimates of the average treatment effects; and (b) the effects continued on the same linear path after 2009. This yields estimates of 0.32 and 0.36 standard deviations, depending on the matching method. We average these to obtain the second lower bound of 0.34.

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 relative student demographics, the school-level focus of test-based accountability (which implies school-level matching maybe preferable), and the fact that the pooled results are less subject to downward bias from trauma/disruption. This upper bound is 0.40 standard deviations. The overall improvement of 0.35 standard deviations (see Figures 1A-1H) is in the middle of this range.[49] We have consulted critics of the reforms and are aware of no alternative theory beyond those, such as population change, that we have already examined and largely rejected.[50]

[48] The difference between panel and pooled appears very small at first, but grows from 2007 through 2009. However, if we assumed the bias continued to grow at the same rate, then the resulting ATE lower bound for 2012 would be smaller than even the unadjusted 2009 panel estimates. This is implausible therefore we use the 2009 bias estimate of 0.23-0.12=0.11 standard deviations.

[49] One additional assumption is that the effects in elementary and middle school do not extend to high school, which we cannot observe. If there are positive treatment effects in high school and those effects accumulate over time, then this assumption makes even our upper bound estimates conservative.

[50] 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.

The above bounds calculations are summarized in Table 7. We also provide a cost-benefit analysis based on the prior frameworks of Krueger (2003) and Harris (2009), using estimates of the reform costs and labor market returns to cognitive skill measured by test scores. Funding levels increased considerably after the storm. The difference-in-difference in operating expenditure between New Orleans and comparison districts was $1,000 per pupil for the year 2009 and onwards (Buerger, 2015).[51] Combining these costs with the effects and implied labor market returns, even the lower bound effects are ten times larger than the break-even value (i.e., where the net benefits equal zero) and much larger than commonly discussed policy alternatives, such as reducing class size and increasing access to pre-kindergarten education.

Effect Heterogeneity

One of the most common critiques of the New Orleans school reforms is that they have been inequitable and even harmful to disadvantaged students. Numerous media reports and lawsuits have alleged denied admission, disproportionate suspensions and expulsions, and insufficient services among certain disadvantaged students under the city's reforms (*P.B. v. Pastorek*, 2010; Jabbar, 2015).

We therefore carried out the same basic estimation methods as above, but separately by FRPL and race/ethnicity. The earlier matching process was modified to add stratification by subgroup.[52] In both the panel and pooled cases, we also carried out many

[51] The $1.8 billion investment in buildings was slow to yield actual improvements in buildings and could not have had a significant influence on these results.

[52] 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.

of the same robustness and bias checks for each subgroup. In general, the sub-group analyses pass these 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 (panel and pooled together). We include only math and language arts for simplicity, though the results are similar for science and social studies. Identification of effects for English Language Learners (ELL) and special education students is left for future research due to several additional methodological issues.[53]

The effects are positive and significantly different from zero for every racial and income subgroup shown.[54] 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.[55] In the panel analysis, black and FRPL students have lower initial effects, but this is followed by similar upward trajectories. This is also true for blacks in the pooled analysis. 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

[53] The ELL population in New Orleans was small before the storm and grew considerably afterwards. Also, there are extremely few ELL students in the comparison districts with which to match. The empirical challenges with special education are a bit different. After the storm, many special education students began taking new types of alternative assessments. There is no crosswalk between these and the regular state tests and the percentage of students taking the alternative assessments changed widely over time. Moreover, there are good reasons to believe that selection into special education worked differently before and after the reforms, which limits us to panel analysis over just the first few years. For these reasons, we leave the analysis of this important topic to a separate study.

[54] These figures also show that the effects usually pass a parallel trends test for each racial subgroup, though not always for FRPL subgroups. 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 but only moderately well in the pooled; specifically, New Orleans white students' pre-Katrina scores in the pooled analysis are considerably above their comparison group means, while New Orleans' black students are below the comparison group.

[55] 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 (multiple comparisons problem).

made them eligible for FRPL.[56] The differences between groups are similar, though not always statistically significant, when we examine other subgroups using panel methods (available upon request).

As before, trauma/disruption effects could explain the discrepancy in the initial dip. 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).[57] Perhaps not coincidentally, these same families also experienced worse initial psychological effects (Brown et al. 2011; DeSalvo, et al., 2007; Elliott & Pais, 2006).[58] We also considered whether the dip for disadvantaged students might have been due to disproportionately low-performing interim schools, but our results are inconsistent with that theory.[59]

[56] 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.

[57] 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.

[58] 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).

[59] 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 larger gains on this crude measure of school quality.

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. Again, it is difficult to distinguish between the trauma/disruption and system effectiveness hypotheses, however.

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. Recall that we include grade repetition as a covariate, and adding this has a minimal effect on the ATE estimates.[60]

While these results are exploratory and there are some inconsistencies, two 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.

Overall, the effect on inequality depends on how we define it. The reforms reduced inequality between advantaged and disadvantaged students statewide by increasing the ranking of this high-poverty, high-minority district from the bottom of the state to nearly the median. But is also increased inequality within the district.

[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.

Additional Evidence

Strategic behavior from test-based accountability remains perhaps the most plausible remaining source of bias because it is hard to test for the resulting test distortions 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. In other work, we have also found no evidence of disproportionate test scores gains near performance thresholds (known as "bubble effects") in New Orleans compared with the rest of the state (Harris, Santillano, & Valant, 2015). However, there is recent evidence of cheating in one New Orleans high school (Dreilinger, 2016).

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. This is what we find: (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) 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); this may be because the choice system leads to better initial matching of students to schools, reducing the need to switch schools (Harris, Valant, & Gross 2015). Racial and income-based segregation has been unaffected (Barrett, Weixler & Harris, 2015), though there are signs that low-scoring students are more concentrated in certain schools (Barrett, Weixler & Harris, 2015) and that low-income students are less likely to choose

[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.

or move to schools that have high test levels (Harris & Larsen, 2015; Maroulis, Santillano, Jabbar, & Harris, 2015). These findings are consistent with the increase within-district achievement gaps and effect heterogeneity reported in Figure 4. Collectively, these other findings are consistent with the effects on achievement.

Conclusion

New Orleans is the first U.S. school system to adopt and sustain an intensive accountability and school autonomy. We find that that the reform package put in place after Hurricane Katrina increased student achievement by a minimum of 0.2, and more likely 0.3-0.4, standard deviations. This means the substantial improvement New Orleans experienced relative to the state was due mostly to the reforms.

The conclusion that the reforms had a positive impact is made possible by a combination of the apparently large magnitude of the reform effects, the sudden and intense nature of the reforms, and the modest magnitude of the potential biases. 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 almost 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 larger 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. It is naturally more likely that such reforms will have similar effects in locations that have similar conditions.

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, p.1209) 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.

Abdulkadiroğlu, A., Angrist, J.D., Hull, P.D., & Pathak, P.A. (2015). School vouchers and student achievement: First-year evidence from the Louisiana Scholarship Program. *Working Paper 2015.06*. School Effective and Inequality Initiative. Cambridge, MA: Massachusetts Institute of Technology.

Angrist, J.D., Abdulkadiroğlu, 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* 126: 699–748.

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 New Orleans 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.

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.

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.

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?* New Orleans, LA.

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.

Dreilinger, D. (2014). Arne Duncan: New Orleans education community is 'leading the nation where we have to. *Times-Picayune*. Downloaded January 1, 2016 from: go'http://www.nola.com/education/index.ssf/2014/12/arne_duncan_new_orleans_educat.html.

Dreilinger, D. (2016). New Orleans high school Landry-Walker's sky-high test scores plunged after cheating probe. Times-Picayune. Downloaded February 14, 2106: http://www.nola.com/education/index.ssf/2016/02/landry-walker_cheating_investi.html.

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 characteristics 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* 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). *Capitalism and Freedom* Chicago: University of Chicago Press.

Fryer, R.G. (2014). Injecting charter school best practices into traditional public schools: 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.

Goldin, C. & Katz, L. (2008). *The Race Between Education and Technology*. Cambridge, MA: 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., Santillano, R., & Valant, J. (2015). Distribution Distortions from Test-Based Accountability in a Market Setting. *Unpublished manuscript*.

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. (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*.

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 Center on Education Statistics (2014). Digest of Education Statistics, Table 216.30. Downloaded January 11, 2016 from: http://nces.ed.gov/programs/digest/d14/tables/dt14_216.30.asp.

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

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.

Plyer, A., Shrinath, N., & Mack, V. (2015). *The New Orleans Index at Ten*. New Orleans LA: The Data Center.

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

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*. Washington, DC.

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.

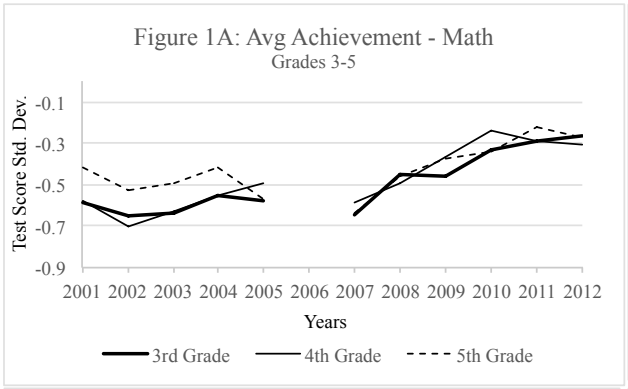
Weems, C. F., Taylor, L. K., Cannon, M. F.,Marino, R. C., Romano, D.M., Scott, B. G., & Triplett, V. (2010). Post traumatic stress, context, and the lingering effects of the Hurricane Katrina disaster among ethnic minority youth. *Journal of Abnormal Child Psychology* 38: 49–56.

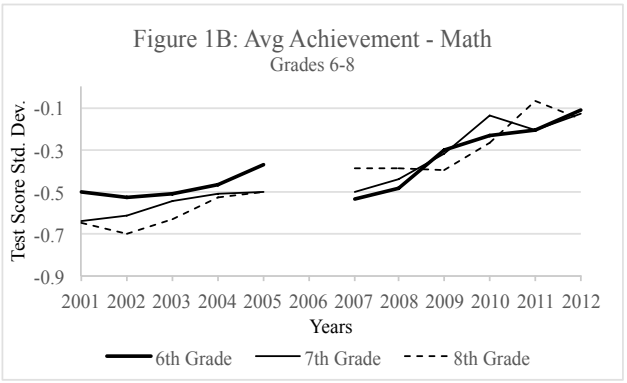
Weixler, L.B., Barrett, N., Jennings, J., Zimmer, R., & Harris, D. (2015). Has the Switch to a Choice System Changed the Distribution of Students by Race, Income, Achievement, and Special Needs Status? Presentation at The Urban Education Future? Lessons from New Orleans 10 Years After Hurricane Katrina.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

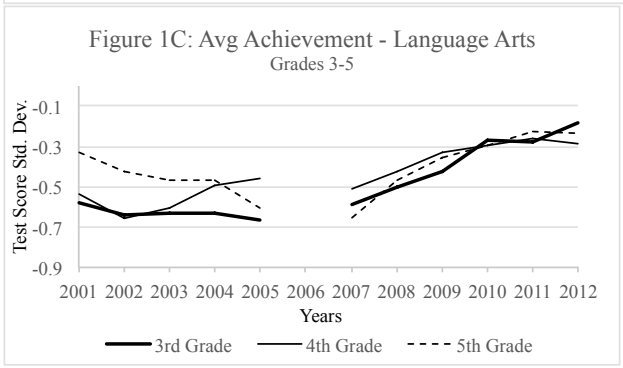
Whitehurst, R. (2012). *The Education Choice and Competition Index: Background and Results 2012*. Washington, DC: Brookings Institution.

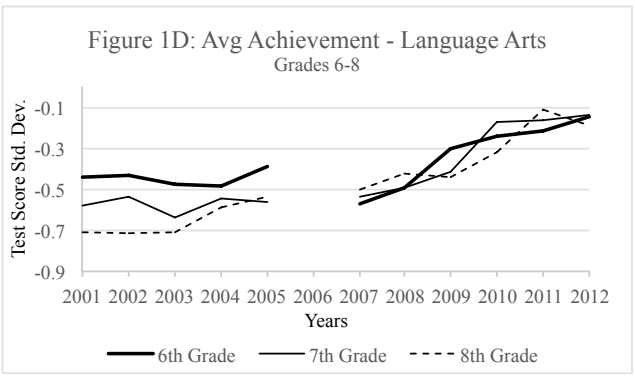
Wong, K. & Shen, F. (2006). Mayors Improving Student Achievement: Evidence from a National Achievement and Governance Database. Paper prepared for the 2006 annual meeting of the Midwestern Political Science Association, April 20-23, 2006.

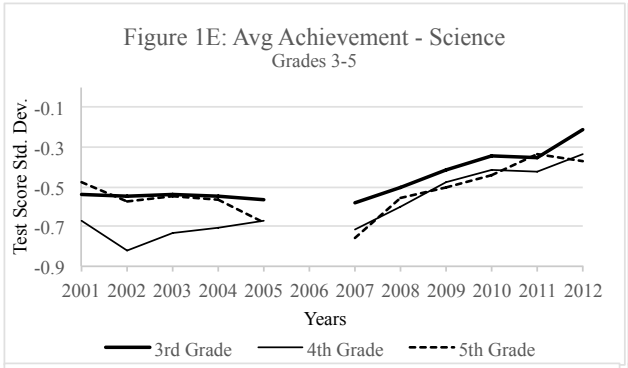
Figure 1: Trends in New Orleans' Student Achievement Levels

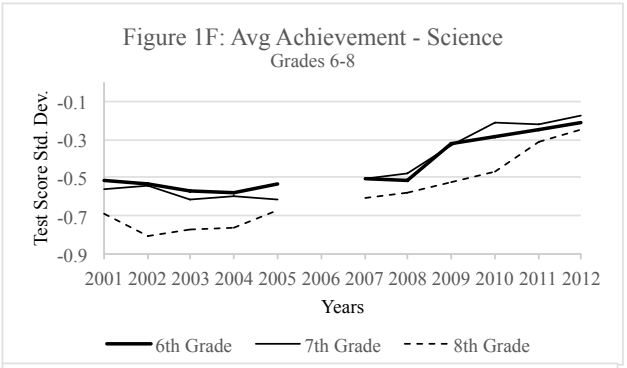


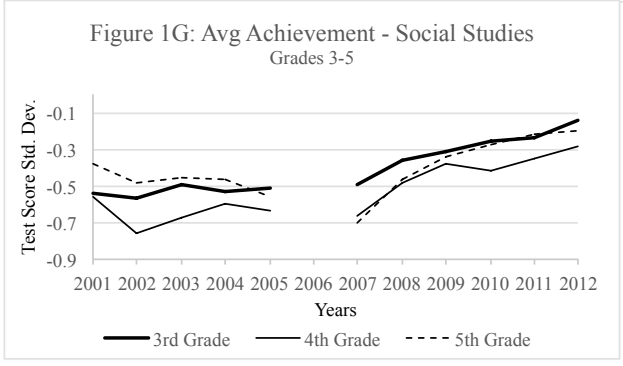


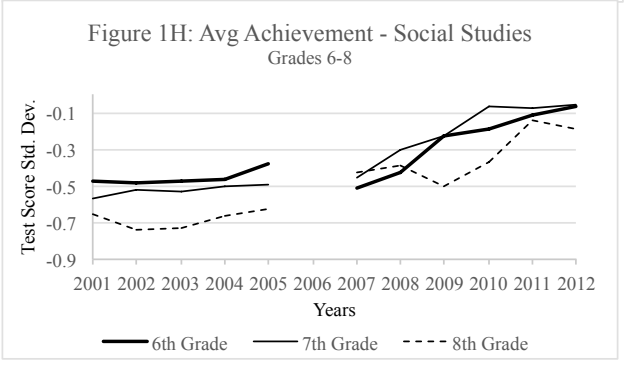








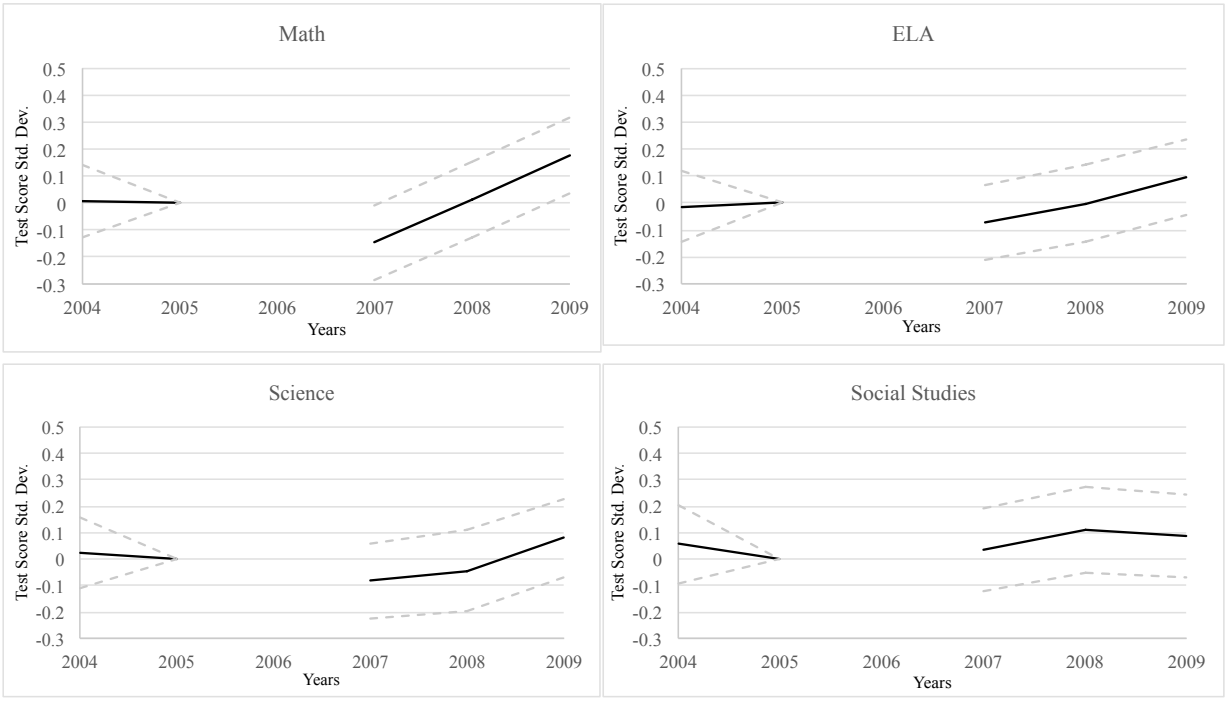




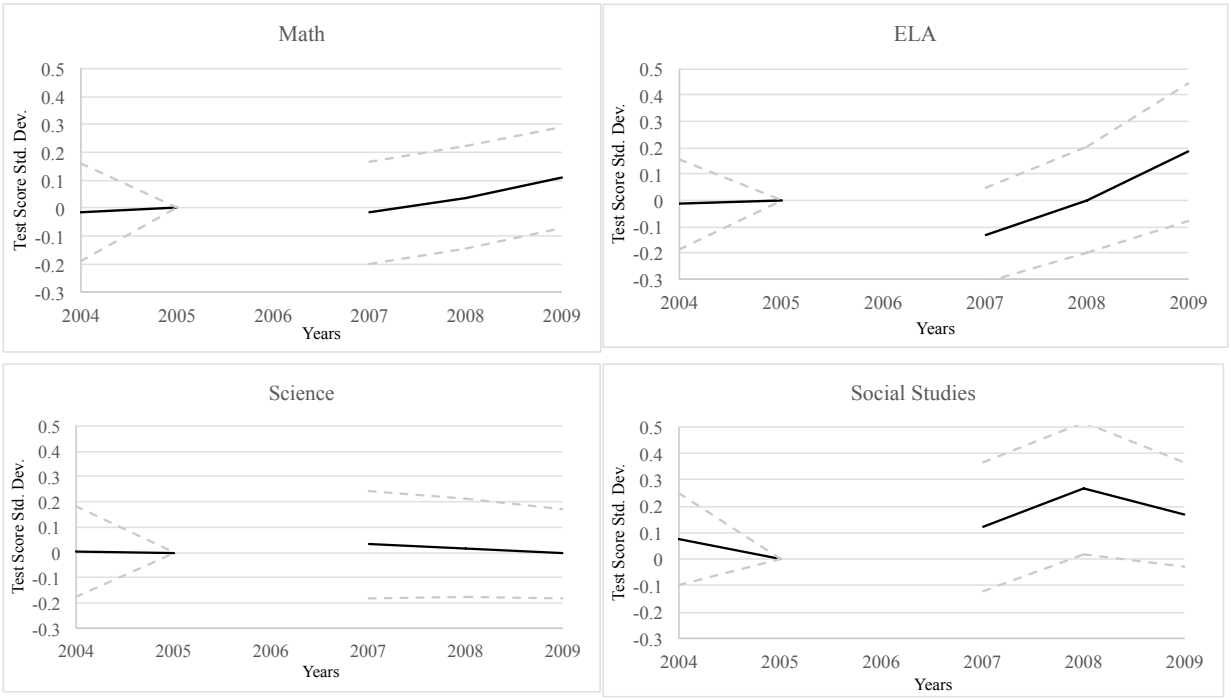
Notes: The y-axis indicates New Orleans test scores standardized to a statewide $\mu = 0$ and σ=1. As the axis suggests, the New Orleans average was below the statewide average in every year, grade, and subject. The break in the middle of the trend lines indicates that the 2005 scores are the last set before the hurricanes and the 2007 scores are the first available in New Orleans post-hurricane.

Figures 2: Panel Estimates of Average Treatment Effects

2005 4th Graders Who Returned in 2006

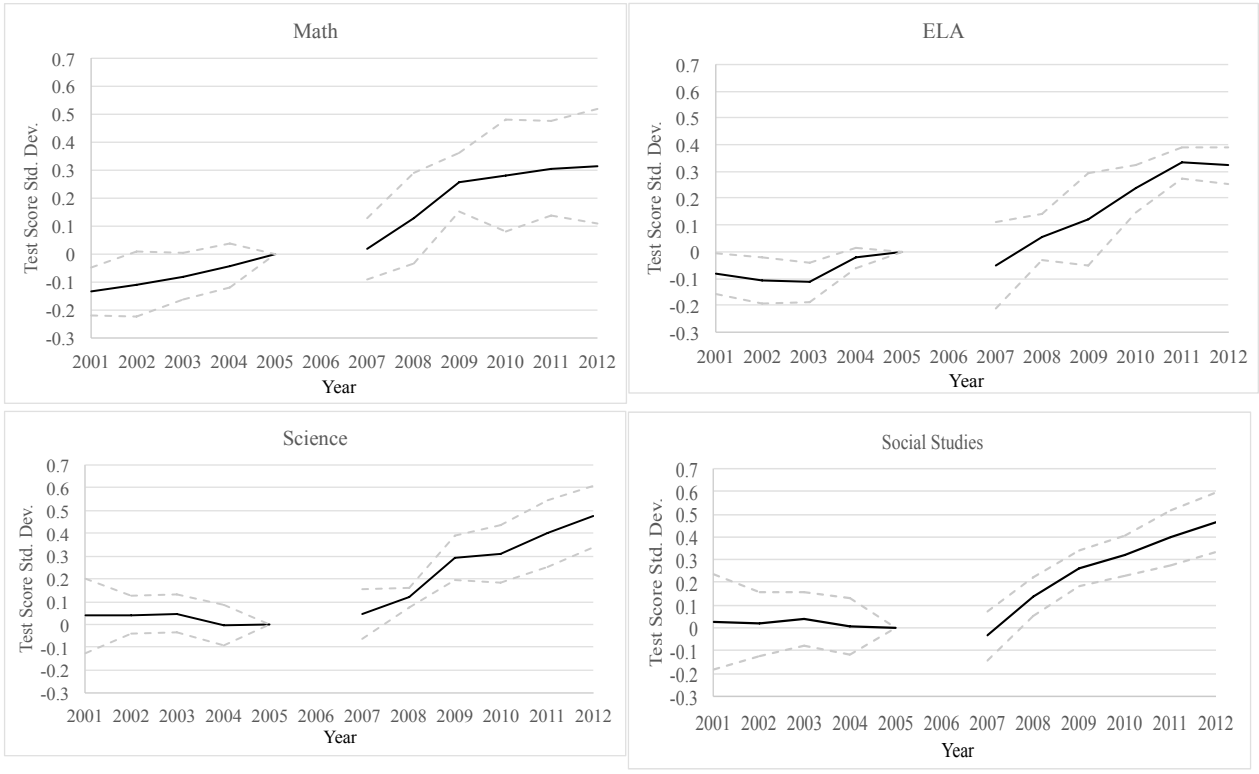


2005 4th Graders Who Returned in 2007



Notes: Results are based on panel estimation of equation (2) using preferred matching on test scores only but without covariate adjustment in the effect estimation. See additional detail in Table 4.

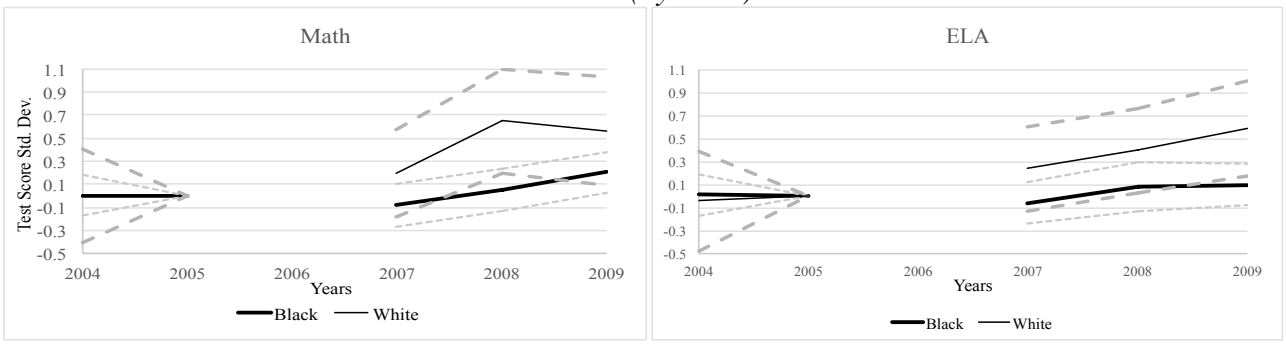
Figure 3: Pooled Estimates of Average Treatment Effects



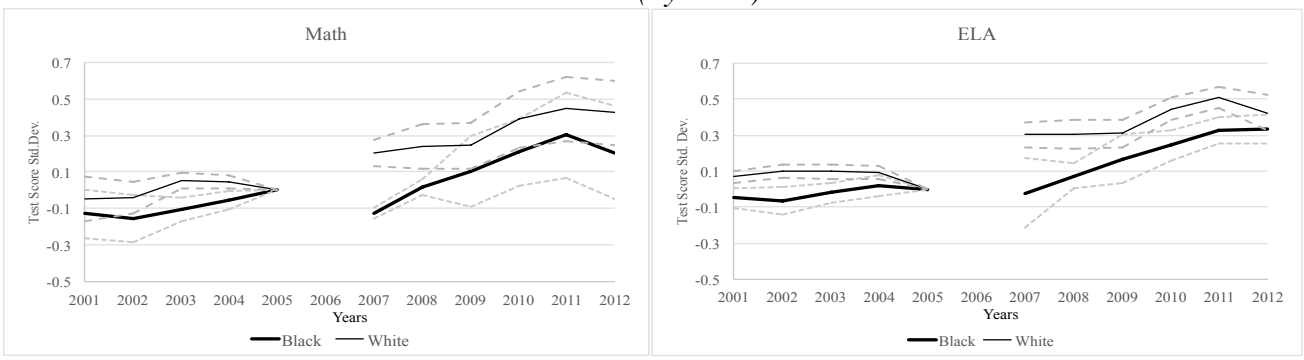
Notes: Effects are averaged across grade levels (weighted). Since these are based on pooled cohorts, and some students are new to the district, they cannot be reported by year of return as they are in Figure 2. See Table 5 for additional details.

Figures 4: Effect Heterogeneity from Panel Analysis

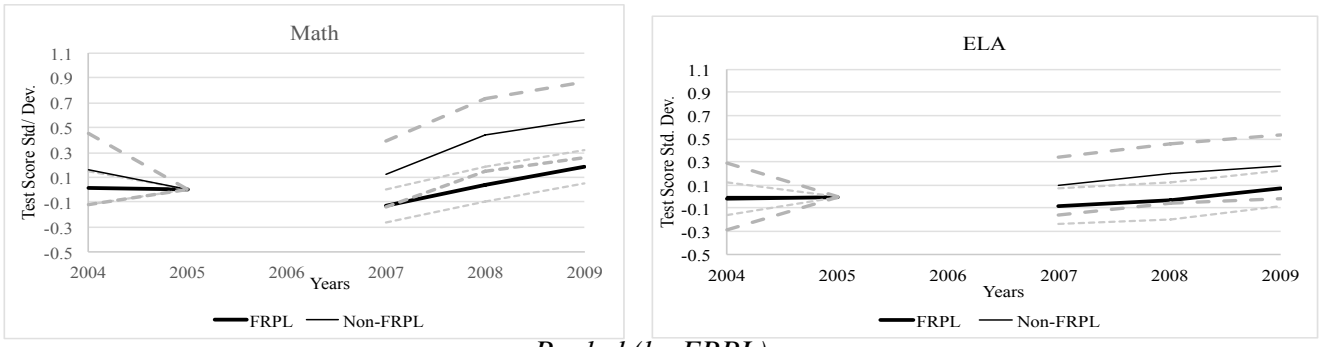
Panel (by Race)



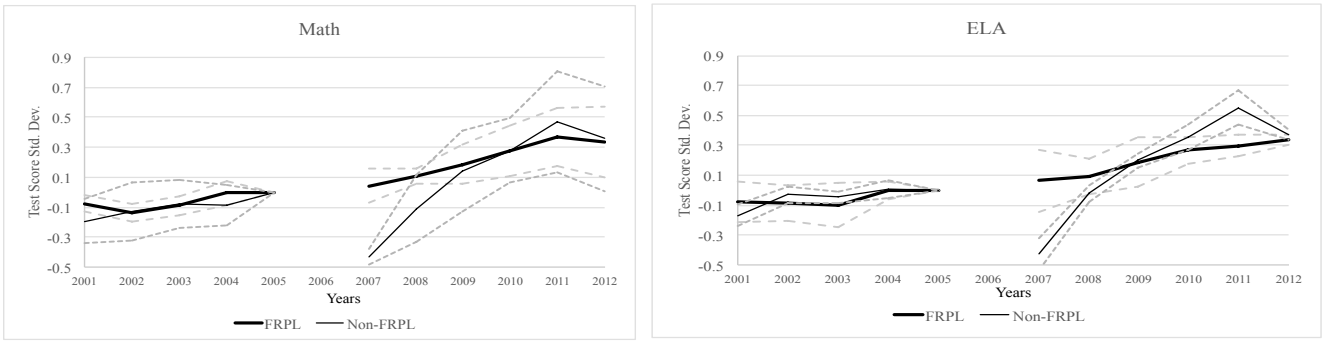
Pooled (by Race)



Panel (by FRPL)



Pooled (by FRPL)



Notes: The panel results are a variation of our preferred panel method where the comparison group is stratified on the subgroup rather than matched. We report here only 4th grade 2006 returnees for the panel. Dashed lines indicate 95% confidence intervals.

Table 1:
Descriptive Statistics for New Orleans Before and After Katrina

Panel A: Demographics		2004-05					2011-12					Mean
		N	Mean	s.d.	Min	Max	N	Mean	s.d.	Min	Max	Diff.
African-American		30,604	0.932	0.252	0	1	18,574	0.895	0.306	0	1	-0.036
Hispanic		30,604	0.012	0.109	0	1	18,574	0.026	0.160	0	1	0.014
Other		30,604	0.021	0.143	0	1	18,574	0.026	0.158	0	1	0.005
White		30,604	0.035	0.185	0	1	18,574	0.052	0.223	0	1	0.017
FRL		30,032	0.841	0.366	0	1	18,492	0.866	0.341	0	1	0.025
Special Education		30,617	0.113	0.316	0	1	18,573	0.099	0.299	0	1	-0.014
ELL		30,617	0.018	0.132	0	1	18,589	0.020	0.142	0	1	0.003
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	4,405	-0.574	0.988	-3.117	3.118	3,118	-0.263	1.022	-3.496	3.043	0.311
Math	4th	6,200	-0.496	1.094	-4.087	3.249	3,340	-0.305	1.020	-4.223	2.656	0.191
Math	5th	4,666	-0.566	0.929	-3.441	2.868	2,715	-0.269	1.038	-3.354	2.917	0.297
Math	6th	4,515	-0.373	0.955	-2.390	3.032	2,931	-0.111	1.066	-3.344	3.102	0.262
Math	7th	4,976	-0.502	0.978	-2.620	2.910	2,721	-0.133	1.101	-3.419	2.699	0.369
Math	8th	5,669	-0.502	1.180	-4.519	2.903	2,832	-0.154	1.125	-4.955	3.588	0.348
Reading	3rd	4,396	-0.665	0.955	-2.911	2.728	3,120	-0.179	1.072	-3.423	3.339	0.486
Reading	4th	6,204	-0.461	1.089	-3.978	3.313	3,338	-0.286	1.116	-4.219	3.164	0.175
Reading	5th	4,670	-0.609	0.943	-3.060	2.507	2,716	-0.231	1.064	-4.152	3.172	0.378
Reading	6th	4,516	-0.391	0.944	-2.294	2.778	2,931	-0.143	1.031	-3.959	3.889	0.247
Reading	7th	4,973	-0.563	0.954	-2.356	2.712	2,726	-0.139	1.028	-3.933	3.116	0.424
Reading	8th	5,370	-0.537	1.127	-4.506	3.104	2,837	-0.190	1.124	-4.862	3.664	0.347
Science	3rd	4,397	-0.568	0.873	-2.942	3.682	3,108	-0.212	1.014	-4.259	3.859	0.356
Science	4th	6,177	-0.669	1.082	-4.213	3.536	3,319	-0.339	1.020	-4.164	3.083	0.330
Science	5th	4,665	-0.678	0.789	-3.080	2.493	2,713	-0.374	1.084	-4.475	4.221	0.304
Science	6th	4,512	-0.528	0.795	-2.446	2.889	2,935	-0.212	1.014	-4.295	3.963	0.316
Science	7th	4,963	-0.614	0.854	-2.665	2.691	2,720	-0.172	1.040	-4.535	3.878	0.442
Science	8th	4,840	-0.670	1.071	-4.035	2.924	2,708	-0.244	1.054	-4.611	3.902	0.425
Social Studies	3rd	4,400	-0.512	0.976	-3.615	2.875	3,107	-0.146	1.027	-3.885	3.846	0.366
Social Studies	4th	6,170	-0.631	1.221	-4.219	2.895	3,319	-0.283	1.090	-4.571	3.903	0.348
Social Studies	5th	4,666	-0.563	0.914	-3.284	3.138	2,716	-0.196	1.086	-4.283	2.985	0.368
Social Studies	6th	4,513	-0.375	0.905	-2.946	3.504	2,934	-0.063	1.038	-4.087	3.906	0.312
Social Studies	7th	4,966	-0.497	0.871	-3.040	3.099	2,723	-0.061	1.056	-4.310	4.077	0.436
Social Studies	8th	4,808	-0.624	1.133	-3.697	3.796	2,704	-0.186	1.116	-4.319	3.769	0.438

Notes: Table 1 includes New Orleans students in the spring testing file for the given year, excluding student who took alternative assessments. The distribution of individual student scores is normalized to statewide $\mu = 0$ and σ=1 for the statewide population within year, grade, and subject. The mean differences in the far right-hand column indicate changes before and after the reforms in the New Orleans population and scores.

Table 2:
Pre-Katrina Mean Differences Between New Orleans and Comparison Group

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		0.920	0.935	0.419	0.664	0.501	0.271
Hispanic		0.011	0.012	0.022	0.019	-0.011	-0.007
Other		0.028	0.020	0.043	0.027	-0.014	-0.007
White		0.041	0.033	0.517	0.290	-0.476	-0.256
FRL		0.864	0.841	0.732	0.785	0.132	0.056
Special Education		0.114	0.114	0.259	0.163	-0.145	-0.050
ELL		0.027	0.018	0.009	0.011	0.017	0.007
Panel B: Test Scores							
Math	3rd		-0.574		-0.131		-0.443
Math	4th	-0.503	-0.436	-0.323	-0.106	-0.180	-0.330
Math	5th	-0.532	-0.561	-0.375	-0.143	-0.157	-0.418
Math	6th		-0.368		-0.298		-0.070
Math	7th		-0.498		-0.345		-0.153
Math	8th		-0.529		-0.274		-0.255
Reading	3rd		-0.663		-0.066		-0.597
Reading	4th	-0.481	-0.428	-0.316	-0.027	-0.165	-0.400
Reading	5th	-0.578	-0.606	-0.414	-0.190	-0.164	-0.416
Reading	6th		-0.388		-0.322		-0.066
Reading	7th		-0.559		-0.319		-0.239
Reading	8th		-0.584		-0.423		-0.161
Science	3rd		-0.567		-0.066		-0.501
Science	4th	-0.623	-0.671	-0.471	-0.021	-0.152	-0.650
Science	5th	-0.668	-0.676	-0.528	0.032	-0.140	-0.708
Science	6th		-0.527		-0.354		-0.172
Science	7th		-0.612		-0.308		-0.304
Science	8th		-0.664		-0.425		-0.239
Social Studies	3rd		-0.512		-0.007		-0.505
Social Studies	4th	-0.528	-0.632	-0.389	0.077	-0.138	-0.709
Social Studies	5th	-0.573	-0.560	-0.452	-0.008	-0.121	-0.552
Social Studies	6th		-0.374		-0.301		-0.073
Social Studies	7th		-0.496		-0.335		-0.161
Social Studies	8th		-0.620		-0.432		-0.187

Notes: All data are from 2005, the school year just prior to Hurricane Katrina. The "Panel" sample only includes those students in 4th and 5th grade who eventually return to their 2005 school district after the hurricane. We report only 4^{th} and 5^{th} grade scores in the panel column because grades 6-8 are not available pre-Katrina. The "Pooled" sample includes all students in tested grades. See details in the text about the matching methods. Most of the differences in the right-hand column are statistically significant at $p<0.05$ and almost are significant at $p<0.10$.

Table 3: Effects of Population Change

Panel A: Population Change (Average Pre-Katrina Characteristics of 3rd Graders)

	New Orleans			Hurricane-Affected Districts			
	All Pre-Katrina Students	Returnees	Diff	All Pre-Katrina Students	Returnees	Diff	Diff-in-Diff
FRL	0.866	0.874	0.008	0.610	0.606	-0.004	0.012
Special Ed	0.101	0.103	0.002	0.164	0.171	0.007	-0.005
ELL	0.017	0.016	0.000	0.034	0.032	-0.001	0.001
Reading Scores	-0.665	-0.683	-0.018	0.118	0.143	0.025	-0.043

Panel B. Census Demographic Changes (Public School Students Only)

	New Orleans			Hurricane-Affected Districts			
	1999	2013	Change	1999	2013	Change	Diff-in-Diff
Income (2013 $)	$43,189	$42,453	-$736	$69,659	$71,408	$1,749	-$2,485
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 difference-in-difference (DD) of demographics and test scores (from LDOE administrative data) between all public school students in 2005 in the respective districts and the returnees in those same districts. Panel B shows DD in district-wide demographics based on Census data (public school students only). Panel C reports regression coefficients based on the federal ECLS, using the same demographics as in the Census; we regressed reading score levels (and gains, separately) 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 carried out an out-of-sample prediction, inserting the Census-based DD changes from Panel B into the regression model in Panel C. The "Cumulative" effects come from adding the effect on 3rd grade test levels to the 5th grade gains multiplied by the dosage (4.2 years per student) to obtain the total predicted effect of demographic change in student test scores. Standard errors of prediction are available upon request.

Table 4A:
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	0.233***	0.166***	0.194***	0.168**
s.e.	(0.055)	(0.059)	(0.057)	(0.071)
Parallel Trends Test	[0.103**]	[0.008]	[0.182***]	[-0.006]
ELA				
Post x NOLA	0.136**	0.100*	0.149***	0.090
	(0.057)	(0.060)	(0.058)	(0.072)
	[0.239***]	[0.016]	[0.208***]	[0.014]
Science				
Post x NOLA	0.232***	0.083	0.218***	0.074
	(0.056)	(0.059)	(0.058)	(0.075)
	[-0.005]	[-0.011]	[-0.006]	[-0.021]
Social Studies				
Post x NOLA	0.243***	0.068	0.258***	0.086
	(0.060)	(0.064)	(0.062)	(0.079)
	[-0.015]	[-0.022]	[-0.036]	[-0.056]
Number of Districts	79	77	8	8
2005 5th Grade Cohort 2005 vs 2008 Diff-in-Diff				
Math				
Post x NOLA	0.167***	0.066	0.166***	0.064
	(0.057)	(0.060)	(0.059)	(0.072)
	[-0.072]	[-0.002]	[-0.109**]	[0.003]
ELA				
Post x NOLA	0.226***	0.027	0.182***	-0.008
	(0.056)	(0.060)	(0.058)	(0.072)
	[-0.258***]	[-0.010]	[-0.229***]	[0.001]
Science				
Post x NOLA	0.092*	-0.026	0.090	-0.114
	(0.054)	(0.058)	(0.056)	(0.071)
	[-0.050]	[0.025]	[-0.094*]	[0.024]
Social Studies				
Post x NOLA	0.217***	0.053	0.201***	0.047
	(0.055)	(0.059)	(0.057)	(0.071)
	[-0.068]	[0.008]	[-0.060]	[0.017]
Number of Districts	78	76	8	8

Table 4B:
Average Treatment Effects from Panel Analysis, 2007 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	0.177***	0.074	0.196***	0.110
s.e.	(0.058)	(0.079)	(0.071)	(0.092)
Parallel Trends Test	[0.075]	[0.076]	[0.061]	[0.016]
ELA				
Post x NOLA	0.156***	0.087	0.160**	0.179
	(0.060)	(0.081)	(0.074)	(0.120)
	[0.155***]	[0.083]	[0.163**]	[0.014]
Science				
Post x NOLA	0.276***	0.097	0.201***	0.006
	(0.059)	(0.082)	(0.071)	(0.089)
	[-0.045]	[0.043]	[-0.040]	[-0.002]
Social Studies				
Post x NOLA	0.310***	0.239***	0.276***	0.182*
	(0.061)	(0.083)	(0.072)	(0.101)
	[-0.099*]	[-0.036]	[-0.131**]	[-0.076]
Number of Districts	67	18	8	6
2005 5th Grade Cohort 2005 vs 2008 Diff-in-Diff				
Math				
Post x NOLA	0.116**	-0.045	0.120*	-0.033
	(0.058)	(0.080)	(0.068)	(0.086)
	[-0.146***]	[-0.025]	[-0.164***]	[0.013]
ELA				
Post x NOLA	0.079	-0.092	0.074	-0.172*
	(0.059)	(0.078)	(0.068)	(0.090)
	[-0.212***]	[-0.036]	[-0.255***]	[-0.020]
Science				
Post x NOLA	0.072	-0.002	0.041	-0.154*
	(0.056)	(0.095)	(0.065)	(0.080)
	[-0.112**]	[0.062]	[-0.160***]	[0.055]
Social Studies				
Post x NOLA	0.210***	0.203*	0.198***	0.025
	(0.059)	(0.121)	(0.069)	(0.086)
	[-0.075]	[0.046]	[-0.120*]	[0.024]
Number of Districts	70	14	8	6

Notes: The first number in each cell is the point estimate for δ in equation (1) with estimation at the student level. Each cell represents a separate regression with robust standard errors (the usual clustering at the district level yielded standard errors that are smaller than the robust one so we report the latter as the more conservative). The results are similar using district-level aggregation. The top portion of each panel pertains to pre-Katrina 4th grade returnees and the bottom portion pertains to pre-Katrina 5th grade returnees. Pre-Katrina 3rd graders are omitted so that parallel trends can be tested (based on 2004 and 2005 test changes). Coefficients and significance levels from the parallel trends tests are shown in [brackets]. See text for discussion of the matching process. *** Significant 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	0.338***	0.368***	0.278***	0.369**
s.e.	(0.027)	(0.045)	(0.059)	(0.117)
Parallel Trends Test	[0.015***]	[-0.015*]	[0.024***]	[0.001]
4th Grade	0.215***	0.327***	0.126***	0.315***
	(0.025)	(0.059)	(0.034)	(0.061)
	[0.041***]	[-0.008]	[0.062***]	[0.050**]
5th Grade	0.281***	0.286***	0.196*	0.230
	(0.030)	(0.034)	(0.088)	(0.187)
	[-0.018***]	[-0.036***]	[-0.006]	[-0.039]
6th Grade	0.261***	0.220***	0.217**	0.156
	(0.024)	(0.037)	(0.066)	(0.122)
	[0.038***]	[0.031***]	[0.048***]	[0.025]
7th Grade	0.402***	0.432***	0.360***	0.392*
	(0.022)	(0.033)	(0.033)	(0.189)
	[0.046***]	[0.023***]	[0.051***]	[0.042]
8th Grade	0.377***	0.460***	0.311***	0.301***
	(0.022)	(0.041)	(0.048)	(0.057)
	[0.062***]	[0.026]	[0.079***]	[0.102***]
Combined	0.312***	0.365***	0.246***	0.304***
	(0.023)	(0.026)	(0.051)	(0.078)
	[0.035***]	[0.004]	[0.048***]	[0.033**]
Number of Districts	109	56	8	7
ELA (Post x NOLA)				
3rd Grade	0.536***	0.539***	0.475***	0.424***
	(0.028)	(0.051)	(0.039)	(0.047)
	[-0.018***]	[-0.036***]	[-0.018**]	[0.032]
4th Grade	0.197***	0.202***	0.144***	0.104**
	(0.020)	(0.043)	(0.037)	(0.034)
	[0.040***]	[0.007]	[0.054***]	[0.040**]
5th Grade	0.371***	0.433***	0.249***	0.329**
	(0.028)	(0.047)	(0.035)	(0.100)
	[-0.064***]	[-0.086***]	[-0.047***]	[-0.048***]
6th Grade	0.253***	0.320***	0.183***	0.241
	(0.019)	(0.052)	(0.028)	(0.134)
	[0.009**]	[-0.013]	[0.018**]	[-0.000]
7th Grade	0.464***	0.516***	0.412***	0.513***
	(0.017)	(0.039)	(0.019)	(0.083)
	[0.009**]	[-0.016]	[0.008]	[-0.001]
8th Grade	0.385***	0.406***	0.345***	0.277***
	(0.013)	(0.036)	(0.022)	(0.070)
	[0.057***]	[0.051***]	[0.067***]	[0.078***]
Combined	0.364***	0.400***	0.297***	0.311***
	(0.018)	(0.024)	(0.023)	(0.030)
	[0.012***]	[-0.012**]	[0.021***]	[0.025**]
Number of Districts	109	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	0.384***	0.402***	0.348***	0.337
s.e.	(0.021)	(0.049)	(0.062)	(0.195)
Parallel Trends Test	[-0.004]	[-0.017**]	[0.005]	[-0.008]
4th Grade	0.363***	0.469***	0.303***	0.563***
	(0.027)	(0.055)	(0.033)	(0.074)
	[0.020***]	[-0.005]	[0.034***]	[-0.009]
5th Grade	0.300***	0.354***	0.272***	0.495**
	(0.026)	(0.036)	(0.069)	(0.134)
	[-0.043***]	[-0.057***]	[-0.040***]	[-0.087*]
6th Grade	0.325***	0.296***	0.307***	0.348***
	(0.019)	(0.043)	(0.051)	(0.067)
	[-0.008**]	[-0.002]	[-0.000]	[-0.007]
7th Grade	0.490***	0.496***	0.499***	0.600***
	(0.016)	(0.031)	(0.036)	(0.114)
	[-0.014***]	[-0.029***]	[-0.019**]	[-0.026]
8th Grade	0.385***	0.406***	0.345***	0.277***
	(0.013)	(0.036)	(0.022)	(0.070)
	[0.015***]	[0.023**]	[0.018]	[0.036**]
Combined	0.390***	0.434***	0.358***	0.469***
	(0.019)	(0.025)	(0.047)	(0.054)
	[-0.002]	[-0.013***]	[0.003]	[-0.013]
Number of Districts	109	56	8	7
Social Studies (Post x NOLA)				
3rd Grade	0.395***	0.332***	0.352***	0.383
	(0.023)	(0.055)	(0.049)	(0.211)
	[0.013***]	[0.004]	[0.014***]	[-0.006]
4th Grade	0.381***	0.431***	0.334***	0.568***
	(0.026)	(0.059)	(0.032)	(0.089)
	[0.009*]	[-0.027***]	[0.020*]	[-0.026]
5th Grade	0.367***	0.371***	0.317***	0.523***
	(0.031)	(0.048)	(0.090)	(0.078)
	[-0.037***]	[-0.059***]	[-0.021]	[-0.093*]
6th Grade	0.321***	0.242***	0.300***	0.391***
	(0.024)	(0.034)	(0.064)	(0.036)
	[0.026***]	[0.007]	[0.035**]	[-0.005]
7th Grade	0.484***	0.459***	0.472***	0.474***
	(0.021)	(0.055)	(0.059)	(0.100)
	[0.022***]	[0.008]	[0.021**]	[0.019]
8th Grade	0.466***	0.445***	0.430***	0.422***
	(0.018)	(0.043)	(0.029)	(0.050)
	[0.020***]	[0.033***]	[0.016*]	[0.048***]
Combined	0.407***	0.395***	0.370***	0.461***
	(0.021)	(0.030)	(0.050)	(0.053)
	[0.012***]	[-0.006]	[0.017**]	[-0.007]
Number of Districts	109	56	8	7

Notes: The first number in each cell is the point estimate for δ in equation (1) with estimation at the student level. Estimation is at the student level with GEE standard errors clustered at the district level. Only 2005 and 2012 scores are included. The third row in [brackets] is the coefficient from the parallel trends test for whether the slope from 2002-2005 different for New Orleans versus the comparison group. See Figure 3 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	0.076***	-0.006
	(0.024)	(0.024)
	{0.037}	{0.046}
ELA		
Post-Katrina	0.110***	-0.021
	(0.024)	(0.024)
	{0.014}	{0.024}
Science		
Post-Katrina	0.100***	0.029
	(0.025)	(0.026)
	{0.040}	{0.040}
Social Studies		
Post-Katrina	0.120***	0.071**
	(0.027)	(0.028)
	{0.026}	{0.048}

Notes: We regress student-level achievement on lagged achievement, grade fixed effects, and an indicator for whether the switch occurred before or after Katrina (Post-Katrina). 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-switchers and the receiving district levels for out-switchers. The number of observed switches ranges from 4,031 to 4,959. 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	-0.070***	-0.072***
	(0.006)	(0.006)
	{0.014}	{0.014}
Switch Type	-0.061***	-0.132***
	(0.017)	(0.013)
	{0.018}	{0.018}
Switch Type*Post-Katrina	0.137***	0.057**
	(0.025)	(0.025)
	{0.041}	{0.044}
ELA		
Post-Katrina	-0.058***	-0.057***
	(0.006)	(0.006)
	{0.011}	{0.015}
Switch Type	-0.101***	-0.111***
	(0.016)	(0.013)
	{0.025}	{0.024}
Switch Type*Post-Katrina	0.155***	0.038
	(0.024)	(0.025)
	{0.021}	{0.027}
Science		
Post-Katrina	-0.058***	-0.065***
	(0.006)	(0.006)
	(0.018)	(0.016)
Switch Type	-0.181***	-0.189***
	(0.018)	(0.014)
	(0.024)	(0.025)
Switch Type*Post-Katrina	0.148***	0.083***
	(0.025)	(0.027)
	{0.047}	{0.041}
Social Studies		
Post-Katrina	-0.046***	-0.056***
	(0.007)	(0.007)
	{0.021}	{0.016}
Switch Type	-0.143***	-0.219***
	(0.019)	(0.014)
	{0.032}	{0.023}
Switch Type*Post-Katrina	0.155***	0.119***
	(0.027)	(0.029)
	{0.035}	{0.043}

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 (*Swtich Type*Post-Katrina*). Robust standard errors are provided in parentheses. In {brackets} underneath are standard errors clustered at the sending district level for in-switchers and the receeiving district levels for out-switchers. 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.

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	0.35
Threats to Identification				
Population Change				
Pre-Kat Scores of Returnees	0.10	0.06	0.04	-0.06
Census/USDOE Simulation				0.01
Interim Schools/Trauma (Pane et al. 2008)	-0.06			
Effects from Panel DD				
Panel (preferred match - tests only)	-0.03	0.05	0.11	[0.32]
Panel (altern. match - tests and demog.)	-0.02	0.06	0.13	[0.36]
Effects from Pooled DD				
Table 5	0.00	0.11	0.23	0.40
Lower Bound - 1				0.29
Lower Bound - 2				0.34
Upper Bound				0.40
Dosage (Post-Reform Years in NOLA)				4.2 years
Annual Cost/Pupil (Buerger, 2015)				$1,000
Adjusted Effectiveness/Cost Ratio (ECR)				
Lower Bound - 1				2.15
Lower Bound - 2				2.51
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. The "Pre-Kat Score of Returnees" for 2007-2009 is the first difference in Table 3 for New Orleans (no comparison group). See definitions for Lower Bound-1, Lower Bound-2, and Upper Bound in the text. The values in the right-hand column under the “Panel” rows are in [brackets] because they are projections of the earlier figures. The "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.