ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-877

Beyond Prescriptive Reforms: An Examination of North Carolina's Flexible School Restart Program

Lam D. Pham
North Carolina State University

Gage F. Matthews
North Carolina State University

Timothy A. Drake
North Carolina State University

While multiple studies have examined the impact of school turnaround, less is known about reforms under the Every Student Succeeds Act (ESSA). To advance this literature, we examine North Carolina's Restart (NCR) model. NCR aligns with ESSA by giving school leaders increased flexibility. Also, NCR differs from previous turnaround models by repackaging a traditionally sanction-based approach to instead motivate school leaders with increased autonomy. Using comparative interrupted time series models, we find positive NCR effects in math, but not in English Language Arts or on non-test-based student outcomes. Also, nearly a quarter of the positive NCR effect can be explained by decreased teacher and principal turnover. These results provide evidence to support current shifts toward reform models featuring local autonomy.

VERSION: November 2023

Suggested citation: Pham, Lam D., Gage F. Matthews, and Timothy A. Drake. (2023). Beyond Prescriptive Reforms: An Examination of North Carolina's Flexible School Restart Program. (EdWorkingPaper: 23-877). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/kfw8-kz84

Beyond Prescriptive Reforms:

An Examination of North Carolina's Flexible School Restart Program

Lam D. Pham

Gage F. Matthews

Timothy A. Drake

North Carolina State University

Abstract

While multiple studies have examined the impact of school turnaround, less is known about reforms under the Every Student Succeeds Act (ESSA). To advance this literature, we examine North Carolina's Restart (NCR) model. NCR aligns with ESSA by giving school leaders increased flexibility. Also, NCR differs from previous turnaround models by repackaging a traditionally sanction-based approach to instead motivate school leaders with increased autonomy. Using comparative interrupted time series models, we find positive NCR effects in math, but not in English Language Arts or on non-test-based student outcomes. Also, nearly a quarter of the positive NCR effect can be explained by decreased teacher and principal turnover. These results provide evidence to support current shifts toward reform models featuring local autonomy.

Keywords

school reform, school improvement, comparative interrupted time series, low-performing schools

Introduction

Policymakers have invested billions of dollars in school reforms to support chronically low-performing schools through federal initiatives such as Race to the Top (RttT) and School Improvement Grants or SIGs (Dragoset et al., 2017; USDOE, 2009). These whole-school reform efforts share an emphasis on integrated, schoolwide interventions, but specific reform models have evolved over time. Earlier models relied on external, often private-sector, partners to manage reforms (Aladjem et al., 2006; Berends et al., 2002), while more recent turnaround models required prescriptive interventions such as replacing the principal and at least 50 percent of teachers (Malen et al., 2002; Strunk et al., 2016). A rich literature has examined these previous models (Redding & Nguyen, 2020; Schueler et al., 2021), but states and districts are again changing their approach to school reform under the current Every Student Succeeds Act (ESSA). Under ESSA, states and districts are still required to develop improvement plans for chronically low-performing schools, but proponents of ESSA's theory of action maintain a core belief that school leaders should have flexibility from bureaucratic restrictions to tailor interventions to school-specific needs. This high-flexibility approach to school reform is a notable shift away from more prescriptive prior approaches like school turnaround, but its effectiveness is not well understood. We fill this gap in knowledge by examining one such high-flexibility model: North Carolina's Restart (NCR) program.

NCR allows local districts to operate persistently low-performing schools with the "same exemptions from statutes and rules as a charter school" (North Carolina General Statute § 115C-105.37B). North Carolina's restart model differs from the federal, SIG-prescribed restart model because it does not require districts to turn over low-performing schools to external organizations. Instead, districts choose to apply for NCR flexibility in their persistently low-

performing schools by submitting an improvement plan to the State Board of Education. If approved, schools remain within their local district but are given "charter-like" flexibilities to carry out their improvement plans. Schools not showing progress after three years under NCR are subject to review, increased support, and potential loss of their NCR flexibilities. Thus, the NCR theory of action eschews top-down sanctions and instead gives low-performing schools the kind of autonomy usually reserved for high-achieving schools (Woessmann, 2007). Moreover, instead of motivating schools to avoid sanctions that might come with external reforms, low-performing schools are motivated to apply for and remain a part of NCR so that they can continue to have the increased flexibility. We describe this new approach in more detail below but emphasize that NCR exemplifies how ESSA is radically shifting the logic of school reform: e.g., from prescriptive interventions to high flexibility, from replacing school leaders and teachers (which we collectively call educators) to increasing their autonomy, and from motivating avoidance of the model to motivating ongoing involvement.

We ask two research questions:

1) To what extent does the NCR program affect student outcomes, including test scores, attendance, chronic absenteeism, disciplinary actions, and the probability of dropping out and graduating from high school?
2) To what extent do changes in educator composition and characteristics (as a result of gaining NCR status) help mediate the effect of NCR on student outcomes?

To answer these questions, we estimate comparative interrupted time series (CITS) models using statewide administrative data that capture all students attending North Carolina public schools from 2011-12 through 2018-19. Our preferred approach compares NCR schools to a comparison group of schools that were nearly eligible for NCR, which we define as schools that meet only

one of the two criteria needed to be eligible. We also show that our results are robust to alternative comparison groups. We find statistically significant and positive NCR effects in math, but not in English Language Arts (ELA) or any non-test-based outcomes. We also coded elements of all NCR school improvement plans and found that plans emphasizing teacher recruitment were associated with significant positive effects. Moreover, we find that a significant proportion of the overall NCR effect in math can be explained by decreased turnover among both teachers and principals. Together, these results provide evidence that NCR can have positive effects, at least on test scores, and they provide preliminary, but not conclusive, evidence to support the ESSA theory of action linking increased flexibility with educator retention and student achievement.

This paper makes four contributions. First, as one of the first evaluations of a whole-school reform model that fully embraces ESSA's emphasis on flexibility for local leaders, we contribute early empirical evidence to illuminate how this theory of action impacts students. Notably, our study is the first to evaluate an ESSA-aligned reform model that relies on incentives (i.e., increased flexibilities) without any direct sanctions on educators. Second, the school reform literature has primarily focused on test scores (Redding & Nguyen, 2020; Schueler et al., 2021), and we examine a broader set of outcomes to contribute a more holistic understanding of how the NCR approach affects students. Third, by examining all NCR schools' improvement plans, we contribute nuanced insights into specific interventions chosen by school and district leaders (e.g., extending the school day and replacing teachers), and compare differences in how these planned interventions affect student outcomes. Because all NCR schools operate within the same set of policy requirements and all in the same state, our study is an advance over previous literature that primarily compare interventions across different reform policies in different states. Finally,

we use rigorous mediation models to contribute insights on mechanisms that help explain how and why NCR affected student achievement. Overall, this work will help guide ongoing school improvement efforts by shedding light on a reform approach that is gaining prominence nationwide.

Literature Review

Policy investment in whole-school reforms have grown over the last few decades in response to frustrations with piecemeal interventions that failed to address obstacles facing the school as a whole (Tyack & Cuban, 1995). Early models such as New American Schools (Berends et al., 2002) and Comprehensive School Reform (Aladjem et al., 2010) attempted to address schoolwide needs by encouraging partnerships with external management organizations. Although these older models highlighted the advantages of addressing schoolwide organizational needs, they left very little decision-making authority in the hands of educators.

Then, under No Child Left Behind (NCLB) accountability policies, states and districts were incentivized to identify and support their lowest-performing schools (Lee, 2008). Using policy initiatives such as RttT and SIGs, the federal government established four school turnaround models (turnaround, transformation, restart, and closure) as preferred approaches to reform (Dragoset et al., 2017; USDOE, 2009). Turnaround models emphasize prescriptive interventions to rapidly improve student achievement, such as replacing the principal and firing then rehiring less than half of teachers (Herman et al., 2008).

Summarizing the effects of these previous reform models, two recent meta-analyses (Redding & Nguyen, 2020; Schueler et al., 2021) report mixed results and show that the literature often narrowly focused on test scores. Reviewing 35 impact evaluations, Redding and Nguyen (2020) found that turnaround reforms are associated with improved test scores, student

attendance, and graduation rates, but the latter two outcomes have been examined by a much smaller set of studies. In a broader review of whole-school reforms under NCLB, Schueler et al. (2021) find a positive effect of 0.06 standard deviation units (SD) in math but null effects in ELA achievement on high-stakes exams, positive impacts on low-stakes exams in STEM and humanities subjects, and no effect among the few studies examining non-test score outcomes. Notably, both reviews compared effects across different interventions (e.g., replacing teachers versus extending the school day). However, they could only compare interventions across different reform models in different state contexts, rather than examining how different interventions play out within one state's policy environment.

Specifically in North Carolina, several studies have evaluated the state's previous school turnaround initiatives. Under RttT, North Carolina initiated a program called "Turning Around the Lowest Achieving Schools" or TALAS, which implemented the SIG-prescribed turnaround models (Henry et al., 2015). Impact evaluations report mixed effects on student achievement that range from -0.05 to 0.09 SD, depending on the sample and model specification (Heissel & Ladd, 2017; Henry & Guthrie, 2019). After TALAS ended, North Carolina continued its turnaround work under the North Carolina Transformation (NCT) initiative. An evaluation of NCT found no significant effect on student test scores in the first year of implementation and a -0.13 SD effect in year two (Henry & Harbatkin, 2020).

Besides evaluations estimating effects on student achievement, research in the school reform literature has also uncovered mediating mechanisms that help explain why some reform models succeed while others fail to improve school performance (Bryk et al., 2010; Henry et al., 2020; Pham, 2022). In a highly influential study of school reform in Chicago, Bryk and colleagues (2010) identified five essential supports that were present in all schools making

progress: leadership, parent-community ties, professional capacity among teachers, student centered learning environment, and instructional guidance. More recent studies continue to support these five essential supports, with researchers finding that experienced leadership (Dixon et al., 2021) and increased teacher effectiveness (Henry et al., 2020; Pham, 2022) are particularly salient mediators in reform models that successfully improved student achievement.

Responding to both the mixed evidence of effectiveness and growing desire for more local autonomy, the federal government gave states and districts more flexibility under the Obama administration's NCLB waivers. These waivers allowed states to either adopt federal turnaround models or to develop their own reforms. Similar to evaluations of the federal turnaround models, impact analyses of these NCLB-waiver reforms showed mixed results on student achievement, with positive effects documented in Kentucky (Bonilla & Dee, 2017) and null or even negative effects in Michigan, Rhode Island, New York, and Louisiana (Atchison, 2020; Dee & Dizon-Ross, 2019; Dougherty & Weiner, 2017; Hemelt & Jacob, 2020).

Near the end of the Obama administration, Congress formally re-authorized the Elementary and Secondary Education Act as ESSA. Following in the footsteps of the NCLB waivers, ESSA delegated more responsibility for school reform to states and districts in response to the criticism that previous federal efforts were too prescriptive. Thus, ESSA requires districts to identify their chronically low-performing schools, but the way schools are identified can be based on non-test-based measures. ESSA also requires states and districts to create improvement plans but does not mandate specific interventions (Klein, 2016). Direct alignment between NCR and ESSA makes this study an important evaluation of how current school reform priorities (i.e., improvement planning and local flexibility) affect student achievement.

The North Carolina Restart Context

To compete for RttT in 2010, North Carolina adopted into law the four federally prescribed models for school turnaround as defined by the SIG program (North Carolina General Statute §115C-105.37B). The SIG restart model required chronically low-performing schools to close and re-open under the governance of an education management organization (Anrig, 2015), which usually meant converting the school into a charter school. Evaluations of the SIG restart model found mixed evidence of effectiveness across different states (Dragoset et al., 2017), but no school in North Carolina chose to implement the SIG restart/charter conversion model (Granados & Hinchcliffe, 2018). Then, in 2015, the North Carolina Department of Public Instruction (NCDPI) began crafting a new policy for persistently low-performing schools (NCDPI, 2016). This new policy created NCR, replacing the SIG restart model with a novel strategy that gives recurring low-performing schools flexibility from state regulations, similar to the flexibility given to charter schools. NCR specifically differs from the SIG restart model because chronically low-performing schools were not removed from their local district. Also, NCR aligns with more contemporary ESSA reform models because NCR schools do not receive additional resources and can instead change how they allocate existing funds. Thus, NCR is inspired by the same core belief that proponents use to support ESSA's theory of action: that school reform decisions should be made by local leaders. NCR was approved in early 2016 and has since been incorporated into North Carolina's ESSA plan (NCDPI, 2017).

In addition to close alignment with federal policy, NCR adds important theoretical insights to the school reform process by repackaging a traditionally sanction-based policy to instead incentivize school leaders with increased autonomy. Most previous reform initiatives, including the SIG restart model, primarily relied on sanctions or threats of sanctions, such as state takeover or replacing educators (for examples, see Schueler & Bleiberg, 2021; Zimmer et

al., 2017). These previous models motivated low-performing schools to improve so they could avoid having to implement reforms. In contrast, NCR does not impose any sanctions on educators and instead allows school leaders to apply for increased flexibility, a practice more commonly used to reward high performing schools (Woessmann, 2007). Thus, instead of avoiding the model, low-performing schools are motivated to ask for and remain a part of NCR so they can retain their autonomy. In other work, we interviewed school leaders about their experience implementing interventions under NCR and found that they overwhelmingly highlighted increased autonomy as an attractive feature of their school for both them and their teaching staff (Matthews et al., 2022).

Thus, evaluating NCR allows us to add several theoretical insights to the literature on school reform. In particular, although limited in scope to one program in one state, our evaluation offers an opportunity to examine whether a model with no direct sanctions on educators can produce positive effects on school performance. Positive NCR effects would be evidence to potentially support current national shifts away from sanctioned-based reform models. Also, we can statistically test whether NCR schools experienced lower teacher and principal turnover, which would provide evidence that the increased autonomy was indeed attractive to educators.

Only schools designated as a "recurring low-performing school" by NCDPI are eligible for NCR. North Carolina defines a school as low-performing if it receives a performance grade of D or F and a growth score lower than "exceed expected growth." Both the performance grade and growth score are calculated under the state's education value-added assessment system (EVAAS) using a combination of student test scores and graduation rates (Granados, 2017). Schools are designated as *recurring* low-performing if they are identified as low-performing in

two of the last three years. The list of recurring low-performing schools is updated every year, and a school named as recurring low-performing can apply for NCR status by working with district leaders and their local school board to submit an improvement plan to the State Board of Education for approval. We note that while NCR schools can employ flexibility akin to a charter school, they must also follow some regulations for traditional public schools. Notably, NCR schools must continue accepting students in their local enrollment area and must continue providing transportation (Granados & Hinchcliffe, 2018). Because of these limitations, NCDPI describes NCR schools as having "charter-like" flexibility.

NCR's requirement that schools undergo improvement planning is another feature that aligns with ESSA's school reform priorities. These improvement plans require schools to list their goals, describe how they will use NCR flexibilities, and set a budget to achieve identified goals. To examine this planning process, we coded all available NCR applications to identify what interventions schools proposed. Our review of NCR applications found that all schools applying for NCR status were eventually approved. From 2016-17 through 2018-19 (the last year of available data), three cohorts of schools were approved for NCR: six schools in 2016-17 (Cohort 1), 67 in 2017-18 (Cohort 2) and 38 in 2018-19 (Cohort 3).[1] The first NCR cohort was much smaller than cohorts 2 and 3 because the program had been recently approved and not yet well-known in 2016-17. Appendix Table 1 shows that school characteristics are quite similar across the three NCR cohorts, and we note that NCR's structure and operating procedures have not changed since it began in 2016. NCR schools operate on a five year review cycle, and NCDPI can choose to require additional interventions if school performance does not improve after the first three years. While schools can elect to exit NCR at any point, NCDPI can begin intensive support for schools that continue to be low performing after three years and can

deauthorize any school that the department believes is not improving. There is currently no time limit on how long a school can remain part of NCR. As of 2018-19, the 111 NCR schools make up 55% of the recurring low-performing schools in NC that are eligible to apply for NCR, and no NCR school has chosen to exit or been removed from the program.

We coded each NCR application for the 14 prominent intervention features identified by Schueler and colleagues (2021) in their recent meta-analysis of the school reform literature: flexibility in funding, governance change from a traditional local school board to another entity (like the state), change in school manager (such as to a charter management organization), human resource changes (such as changes in teacher pay or evaluation systems), teacher professional development, assistance for administrators, flexibility in teacher hiring or assignment (e.g., hiring uncertified teachers), replacing principals, extended learning time, tutoring, curricular change, data use to inform instruction, wraparound services, and school choice (i.e., although the NCR school must accept all students residing in their local enrollment area, students assigned to an NCR school may choose a different school). For definitions for each intervention feature, as developed by Schueler et al. (2021), see Appendix Table 2. We attempted to code all 14 intervention features, but 2 features (governance change and principal replacements) were not proposed in any NCR plans, and 3 features (human resource changes, teacher professional development, and extended learning time) were so commonly proposed together that we could not report them separately. Thus, we created a "common intervention" indicator that equals one if schools propose any of these three features that almost always appear together. These restrictions left us with 10 intervention features for analysis. Table 1 below lists the number of schools and districts proposing each feature. Among the 111 NCR schools in our analysis, 65 proposed some flexibility from existing procedures for hiring and assigning teachers,

49 proposed a curriculum change, 35 proposed additional support for administrators, 34 proposed changing budget allocations, 4 schools in one district proposed allowing students to enroll in a different school, and 101 proposed the common intervention (either human resource changes, teacher professional development, or extended learning time).[2]

Methods

Data and Measures. We use statewide longitudinal data on all students in North Carolina public schools, collected by NCDPI and managed by the North Carolina Education Research Data Center. These data contain rich student characteristics such as gender, race, and indicators for economically disadvantaged students (ED), multilingual learners (ML), and students with disabilities (SWD). The data span 2011-12 through 2018-19, capturing five years before and three years after NCR began.[3]

These data provide several student outcome measures. For math and ELA test scores, we use the state-required end-of-grade exams in grades 3-8 and end-of-course exams in high school, which we standardize within year, subject, and grade to have a mean of zero and unit variance. For attendance rate, we use the proportion of enrolled instructional days in which the student attended school. Chronic absenteeism is measured with an indicator for students who have an attendance rate at or below 90 percent[4] (NCDPI, 2021). To measure disciplinary actions, we use indicators for whether the student has ever committed a reportable offense[5] and whether the student has ever received a long-term suspension. We chose these two measures because they are more likely to be similarly reported across schools, regardless of individual school or district disciplinary policies. Finally, we use state graduation records to code an indicator for whether the student drops out of school in any year and an indicator for whether the student ever receives a high school diploma.

These data link students with the schools they attend each year, allowing us to identify students who attended NCR schools and to aggregate student demographic characteristics to the school level. We augment these student-level datasets with school-level data from NCDPI capturing schools' EVAAS performance grade and achievement scores. These measures allow us to identify low-performing and recurring low-performing schools, as defined by NCDPI. Also, we include data from the National Center for Education Statistics (NCES) capturing school characteristics, including total enrollment, locale (rural, town, suburb, or city), and school level (elementary, middle, high, or other school level). We also use data from our own coding of NCR applications, which includes the year schools first began implementing their NCR reforms and the intervention features they proposed.

The educator level datasets contain demographic and professional characteristics, and teachers can be linked to students using course membership files (NCERDC, 2009). For both teachers and principals, we generate indicators for movers who transfer schools and leavers who exit the dataset completely. Other educator measures include years of experience, graduate degree attainment, salary in $10,000 dollars, and standardized Praxis exam scores. For teachers, we can also identify whether they are working in their licensed subject and whether they are alternatively certified. We also estimate teacher value-added (VA) scores for ELA and math separately. The VA scores are calculated from regressing student test scores in the current year on math and ELA scores in the prior year with controls for student-, class-, and school-level characteristics. The model also includes grade and year indicators, and teacher-by-year fixed effects that are constrained to sum to zero. To account for measurement error in our value-added measure, we retain only teachers who can be connected to 10 or more students, and we use the

empirical Bayes method to shrink the predicted teacher-by-year fixed effect, which serve as our VA scores.

Sample. Our sample includes schools that ever become an NCR school as the "treatment" group. The strength of our CITS model (described below) depends on identifying a comparison group that is a plausible counterfactual for how schools would have fared in the absence of NCR. Our preferred comparison group comprises schools that are nearly eligible for NCR. Under NCDPI's definition, schools must receive a performance grade of D or F *and* a growth score lower than "exceed expected growth" in two of the previous three years to be eligible for NCR (or any of the other state-approved reform models); therefore, we use the schools that either receive a D/F performance grade *or* a growth score lower than exceed expected growth (but not both) in at least two of the three years between 2013-14 and 2015-16. We consider these schools nearly-eligible for NCR because they meet only one of the two criteria for designation as a recurring low-performing school. This comparison group offers several advantages. First, meeting either of the two criteria is a signal that the school is low-performing in ways that are likely comparable to schools that meet both criteria. Second, being nearly-eligible in two of three years protects against potential bias from mean reversion in schools that posted one year of abnormally low performance. Third, these nearly-eligible schools are not eligible for any of the other reform models available to recurring low-performing schools in North Carolina (i.e., turnaround, transformation, closure), so any effects we find will not be influenced by alternative reform models occurring in the comparison group. Finally, we use only schools that are nearly-eligible based on data from the three years before NCR began. We did not include any schools that become nearly-eligible after 2016-17 because, after becoming aware of NCR, these schools may experience some form of treatment if being nearly-eligible motivates them to implement

alternative improvement efforts. Thus, our preferred sample includes only the 111 schools that ever implement NCR between 2016-17 and 2018-19 and 135 schools that are nearly-eligible based on data from the three years before 2016-17. Our preferred sample does not include any schools that joined the NCR program after 2018-19. For power analyses, please see Appendix Table 3.

Table 2 shows descriptive characteristics of NCR and nearly-eligible comparison schools in the years before and after NCR began. The table also shows results from *t*-tests for significant differences between the two groups of schools. In the baseline years before reforms began, about half (53%) of the students in NCR schools are Black, 22% are Latino/a/x, and 18% are white, compared to 48%, 22%, and 20% in comparison schools. Moreover, about 15% of students in both groups of schools are identified as having disabilities, 75% are economically disadvantaged, and 11% are multilingual learners. None of these student demographic characteristics differ significantly between NCR and comparison schools, and all remain stable after reforms began. The only baseline school characteristics that are significantly different between NCR and comparison schools suggest that comparison schools are more likely to be located in towns and less likely to be in suburban areas than NCR schools. Although statistically significant, we do not see these differences as cause for major concern because towns and suburbs are not starkly different in terms of population density, and the magnitude of the difference is not substantively large (about 9 percentage points). Overall, Table 2 suggests that the low-performing schools in our sample primarily serve low-income and minoritized students, and similar descriptive characteristics support the nearly-eligible schools as a valid comparison group.

We also test whether results are robust to alternative comparison groups. Our first alternative comparison group includes all the schools that are eligible for NCR between 2016-17

and 2018-19 but did not apply for NCR status. These schools are most comparable to NCR schools in terms of having both a low performance grade and growth score, but they may suffer from unobserved selection issues that differentiate them from schools that chose to apply. Second, we use a group of schools that applied for NCR but rescinded their applications before implementing any reforms. Similar to treated NCR schools, rescinded schools were motivated to apply for NCR, but they also suffer from potential selection issues based on why they decided against staying in the program. Our investigation of rescinded schools found multiple reasons for their withdrawal, including new school or district leadership that was not invested in NCR, wanting to focus on other district/school initiatives, and, in one case, a natural disaster that closed the school building. Finally, we test a comparison group of future NCR schools. Our coding of NCR applications revealed 41 schools that applied for NCR after 2018-19. We do not include them in our primary analysis because our administrative data contains only student outcomes through 2018-19; however, we test results using them as comparison schools (i.e., in the years before they implement any NCR reforms). Like treated NCR schools, these future NCR schools were motivated to apply for the program, but there may be unobserved factors related to student outcomes that made them hesitate to apply when NCR first began.

Analytic Models. We estimate the effect of NCR reforms on student outcomes, y, using a CITS model (Bell et al., 2016; Bloom, 2003; Dee & Jacob, 2011; Somers et al., 2013; Steinberg & Sartain, 2015; Strunk et al., 2016). This model compares deviations from baseline trends among NCR schools with analogous deviations in nearly-eligible comparison schools; resting on the identifying assumption that any difficult-to-observe factors influencing student outcomes in the absence of NCR are captured by the comparison schools' deviation from their baseline trend. That is, CITS results are unbiased if confounders are not systematically different between NCR

and comparison schools, so any additional deviation from the baseline trend in NCR schools can be attributed to the NCR reforms.

We estimate the CITS separately for each outcome of interest: standardized test scores in ELA and math, attendance rate, and whether the student is chronically absent, committed a reportable offense, received a long-term suspension, dropped out of school, or received a high school diploma. We estimate the following model for student i, in school s, and year t:

$$y_{ist} = \beta_0 + \beta_1 NCRYear_{st} * EverNCR_s + \alpha_1 Yr1_{st} + \alpha_2 Yr1_{st} * EverNCR_s + \alpha_3 Yr2_{st} + \alpha_4 Yr2_{st} * EverNCR_s + \alpha_5 Yr3_{st} + \alpha_6 Yr3_{st} * EverNCR_s + X'_{ist}\delta + \phi_s + \theta_t + \varepsilon_{ist} \quad (1)$$

Equation 1 regresses the student outcome (y) on the interaction between $EverNCR$, an indicator for whether the student attends a school that ever participates in NCR, and *NCRYear*, a linear trend variable centered at zero in the year before schools receive NCR flexibilities and representing the number of years before and after NCR reforms began.[6] For cohort 1 schools, *NCRYear* is zero in 2015-16, 1 in 2016-17, 2 in 2017-18, and 3 in 2018-19. For cohort 2, *NCRYear* is centered at zero in 2016-17, and for cohort 3, it is centered at zero in 2017-18. *Yr1*, *Yr2*, and *Yr3* are dichotomous variables indicating the first, second, and third year after schools begin NCR reforms. For comparison schools, *Yr1* equals 1 in 2016-17, *Yr2* equals 1 in 2017-18, and *Yr3* equals 1 in 2018-19. Additionally, Equation 1 controls for a set of student characteristics (X_{ist}) including indicators for gender, race, SWD, ED, and ML status. For math and reading outcomes, X_{ist} includes the student's prior year lagged test score. Finally, Equation 1 includes school (ϕ_s) and year (θ_t) fixed effects, with standard errors clustered at the school level. The school fixed effect controls for any time invariant factors affecting both the probability of schools applying for NCR and student outcomes. The year fixed effect controls for any global trends across time in student outcomes and is especially important for math test scores because

North Carolina changed its math standards and adopted new math exams in 2018-19, resulting in a statewide dip in average math performance (NCDPI, 2019). Note that the year fixed effect is more flexible than CITS models that instead specify linear global time trends.

In Equation 1, β_1 is the difference in the pre-NCR (or baseline) trends between NCR and comparison schools. The coefficients on the Yr1-Yr3 indicators (α_1, α_3, and α_5) estimate the average deviation from baseline trend in years 1, 2, and 3 for comparison schools. The coefficients of interest are α_2, α_4, and α_6, representing cumulative deviations from the baseline trend in NCR schools minus the same deviation for comparison schools in years 1, 2, and 3 after NCR reforms are implemented. For all the dichotomous outcomes, Equation 1 is estimated as linear probability models.

The CITS model has been used in prior evaluations of educational policies (Bell et al., 2016; Dee & Jacob, 2011; Steinberg & Sartain, 2015; Strunk et al., 2016) and offers several advantages. First, modeling pre-treatment trends relaxes the parallel trends assumption in a difference-in-differences (DID) framework (Somers et al., 2013).[7] Second, the model allows us to examine both deviations in means and trends across time. Third, flexibility in how the model is specified allows us to test whether results are robust to our modeling choices, including what covariates are included and how trends are modeled.

We also estimate the effect of NCR in each cohort separately using three subsamples, where each subsample retains only comparison schools and one cohort of NCR schools – dropping all NCR schools in earlier or later cohorts. Then, we examine relationships between proposed intervention features and student achievement by separating the EverNCR indicator into a set of indicators for the ten intervention features listed in Table 1. To facilitate interpretation, we collapse the three *Yr1*, *Yr2*, and *Yr3* indicators into one *AfterNCR* indicator

that equals 1 in any year after schools begin implementing NCR reforms. Because intervention features are not randomly assigned, these analyses are descriptive rather than causal. Also, we examine whether certain student characteristics (e.g., gender, race) moderate the impact of NCR reforms using three-way interactions between these student characteristics, and indicators for *EverNCR*, *NCRYear*, and *AfterNCR* (along with all main effects and two-way interactions). We use these analyses to examine whether NCR reforms have heterogeneous effects across student subgroups.

We have several reasons for choosing CITS over newly developed DID approaches that address dynamic treatment effects from staggered treatment timing (Callaway & Sant'Anna, 2020; Goodman-Bacon, 2021). First, we have no evidence suggesting that the structure, implementation, or effect of NCR changed over the three years and cohorts of NCR schools. Schools applied to join using the same procedures and were offered the same flexibilities in all three years. Also, we find no systematic changes in what types of schools joined NCR across the three cohorts (Appendix Table 1). Second, these staggered-treatment-DID models are still under development and cannot yet fully accommodate several crucial components in our preferred model, e.g., school and year fixed effects. Also, our cohort-specific estimates produce similar conclusions to results from the model pooling together all three cohorts. These cohort-specific analyses do not suffer from dynamic treatment effects because NCR schools in the same cohort began reforms at the same time, suggesting that NCR effects are stable across the three years in our analysis.

Finally, to explore mediating mechanisms, we examine how NCR affected the characteristics of educators in NCR schools. We focus on educator characteristics as mediators because previous literature supports teachers and leaders as two of the most impactful reform

mechanisms (Bryk et al., 2010; Dixon et al., 2021; Henry et al., 2020). Specifically, we estimate how gaining NCR flexibilities affected educator turnover (from both moving and leaving), race, experience, degree attainment, salary, and licensure. For teachers, we also examined Praxis scores, whether teachers taught in their licensed subject, and VA scores. To test these mediating mechanisms, we use a standard mediation approach, illustrated in Figure 1 (Baron & Kenny, 1986; Preacher, 2015). Figure 1 depicts four pathways: Path C is the overall or total effect of NCR; Path A is the effect of NCR flexibilities on the mediator; Path B is the relationship between the mediator and outcomes; and Path C' is the direct effect of NCR on outcomes after controlling for observed mediators. In this framework, Path A * B, our main pathway of interest, is the indirect effect of NCR reforms on student outcomes through the mediator. That is, $C = C' + A * B$. For example, to test teacher turnover as a mediator of the NCR effect on math test scores: Path A would be the effect of NCR on teacher turnover; Path B would be the relationship between teacher turnover and math test scores; Path $A * B$ would be the indirect effect of NCR on math test scores through its effect on teacher turnover; and C' would be the direct effect of NCR after controlling for teacher turnover.

To estimate Paths $A, B, and\ C'$ simultaneously, we use structural equation modeling (SEM) to estimate our CITS model with the inclusion all mediators of interest. Note that the CITS model we use for the mediational analysis is similar to Equation 1 but uses one *AfterNCR* indicator instead of separate indicators for Years 1-3. Although these mediational analyses are non-causal, we can test whether the indirect effect is statistically significant using a bootstrap procedure with 5,000 replication samples that randomly draws whole schools with replacement to preserve the clustering of students and educators within schools (Preacher & Hayes, 2004, 2008). We then use the 5,000 replications to derive bias-corrected confidence intervals that test

the statistical significance of each indirect effect estimate (Preacher & Hayes, 2004, 2008). Simulation studies support these bootstrap confidence intervals because they do not make assumptions about the distribution of the indirect effect, but they can be asymmetric (Preacher, 2015).

Results

Table 3 shows results from estimating Equation 1 on math and ELA test scores, with columns 1 and 5 using nearly-eligible schools as our preferred comparison group. (For a descriptive plot of math and ELA trends, see Appendix Figure 1.) Our primary coefficients of interest (α_2, α_4, and α_6 in Equation 1) are shown in the three rows labeled as Ever NCR * Years 1, Ever NCR * Years 2, and Ever NCR * Years 3. Overall, we find positive NCR effects in math, but no significant effects in ELA. By year 2, the cumulative deviation from baseline trends in math for NCR schools is 0.113 SD higher than the same deviation in comparison schools. In year one, the math effect is positive, but smaller in magnitude (0.024 SD) and marginally significant at the 10 percent level. In year 3, the coefficient in math is also positive but not significant. The year 3 estimate may not be significant because only one cohort of NCR schools has three years of post-NCR outcomes, resulting in low statistical power (see Appendix Table 3). Indeed, the year 3 math estimate is similar in magnitude but less precise than the year 2 estimate. Table 3 also shows that results are robust to our choice of comparison group. For ELA, the coefficients are not significant relative to any comparison group. For math, the effect estimates are consistently significant across different comparison groups in year 2. This result in math continues to be statistically significant when we correct for multiple hypothesis tests by adjusting *p*-values using a false discovery rate control method developed by Benjamini et al.(2006), see Appendix Table 4. In contrast, the year 3 estimates are only significant when future NCR schools

are used as the comparison group. Because this one significant estimate in year 3 is not robust across different comparison schools, we emphasize that our results for year three are not conclusive. To test for effect heterogeneity, Appendix Tables 5 and 6 examine the moderating effects of student and school characteristics for math and ELA, respectively. We find no evidence that the NCR effects are significantly different for any of the student or school-level moderators.

When estimating results separately for each cohort (Appendix Table 7), we observe a similar pattern where the estimate increases in math between year 1 and year 2 but does not increase further in year 3. However, none of the cohort-specific estimates are significant in ELA or math. We conclude that NCR flexibilities had a positive effect on student achievement in math and that this effect grows larger after a full year of implementation. Whether the effects are maintained after the first couple of years is an open question that will require additional data to convincingly answer.

Table 4 reproduces results in math and ELA and adds results for attendance rate, chronic absenteeism, reportable offenses, long term suspension, dropout, and graduation. Henceforth, we report only results using nearly-eligible schools as the counterfactual, which we call comparison schools. We find scant evidence of positive or negative effects on any of the non-test score outcomes. Most of the coefficients have a positive valence (i.e., increased attendance rate and probability of graduation and decreased probability of chronic absenteeism, reportable offenses, and long-term suspension); however, the estimates are small in magnitude and not statistically significant, i.e., precisely estimated zeroes.

Table 5 reports results when we replace the *EverNCR* indicator with ten separate indicators for each of the intervention features coded in schools' NCR applications. These results

are descriptive, and we emphasize that they only capture relationships between *planned* interventions and student outcomes. For clarity, Table 5 reports coefficients on the interaction of each intervention feature with an *AfterNCR* indicator (rather than separate indicators for years 1-3) and focuses only on test scores because we did not observe significant effects for any of the other outcomes. Holding constant all other proposed intervention features, we find that changing the curriculum is associated with a 0.058 SD decrease in math scores. None of the other interventions are significant at the five percent level. Furthermore, we estimate a model that replaces the *EverNCR* indicator with a continuous variable that sums up how many intervention features are proposed in the school's NCR application (Appendix Table 8). This model allows us to examine whether schools proposing more interventions produce larger positive effects on student achievement. We find that increasing the number of proposed interventions is not associated with improved math or ELA scores. In fact, we find a significant negative relationship with math scores by year three.

Finally, Figure 2 illustrates results from the mediational analyses on math test scores. We focus only on math test scores because it was the only outcome where we detected consistently significant NCR effects. Note that Figure 2 depicts only the four mediating mechanisms that were statistically significant at the 5 percent level, but for both teacher and principal level characteristics, our models included all mediators at the same time. For full results, see Appendix Table 9. Estimates for Path A in Figure 2 suggest that NCR schools experienced (1) decreases in the probability that teachers will move into another school, (2) increases in teacher salary, (3) increases in the probability that teachers are alternatively licensed, and (4) decreases in the probability that the principal will leave the dataset. Decreasing teacher turnover from moving and principal turnover from leaving partially explains the positive NCR effect because the Path B

estimates show that both teacher movers and principals leavers are associated with lower test scores. In contrast, teacher salary and probability of being alternatively licensed are associated with increases in math test scores.

For each mediator, Figure 2 also displays the indirect, mediating effect (Path A*B), the 95% bias-corrected confidence interval from our bootstrapping procedure, and the percent of the total NCR effect explained by each mediator. For reference, when we estimate the CITS effect averaging together the three post-NCR years, the total NCR effect on math test scores is 0.04 SD. As an example, the results show that the indirect effect of NCR flexibility on math scores through reducing teacher turnover is 0.003 SD, which accounts for about 8% of the total NCR effect (0.003/0.04). Taken together, the four significant mediating mechanisms (reducing teacher and principal turnover and increasing teacher salary and alternative licensure) explain about 40% of the total NCR effect.

Because increasing average teacher effectiveness in a school could occur by either hiring more effective teachers, pushing out low-performing teachers, or developing current teachers, we also estimated the same mediation models on subsamples of incoming and outgoing teachers to better understand how NCR reforms were changing the composition of educators in NCR schools (Appendix Tables 10 and 11). These analyses yielded three statistically significant mechanisms: incoming teachers in NCR schools had higher salaries, lower VA scores, and higher Praxis scores (after controlling for other teacher characteristics like experience).

Robustness and Validity Checks

Besides testing the robustness of our results to different comparison groups, Appendix Table 12 compares our preferred model with a series of alternative specifications where we (1) remove all student covariates; (2) replace the school fixed effect with school characteristics

measured in the baseline (pre-NCR) year; (3) replace the school fixed effect with a district fixed effect; and (4) weight all models by school enrollment. Results across these alternative specifications are largely robust, except the math results are similar in magnitude but less precise and therefore no longer significant in the district fixed effect model. We also observe larger positive effects on math scores in year 3 across some of these alternative specifications, but given the small number of schools with available year 3 results, any additional interpretation of year 3 effects should await future work with additional data.

It is also possible that our results are driven by small-sample bias from an over-parameterized CITS model with artificially low standard errors. To address this issue, Appendix Tables 13-15 show a set of results from more parsimonious models. Appendix Table 13 shows unadjusted mean differences between NCR and comparison schools in the years after NCR began (i.e., a simple linear regression of math and ELA scores on an indicator for NCR schools). Appendix Table 14 shows a simple DID specification with no covariates and a DID model with only student covariates (gender, race, SWD, ED, and ML status). Appendix Table 15 shows a simple CITS model with no school or year fixed effects. Results in all of these simpler models show positive and significant effects in math but not in ELA. Finally, to show that our results over time are also robust to different specifications, Appendix Figure 1 graphs unadjusted mean test scores in each academic year, and Appendix Figure 2 shows estimates from a simple event-study model that regresses test scores on indicators for each year before and after schools began NCR reforms, with the baseline year before NCR began (year 0) as the reference year. Both figures support our preferred results in Table 3, showing no significant differences between NCR and comparison schools in the years before NCR began, positive effects in years 1 and 2 in math,

and null effects in ELA. Together, these descriptive averages and simplified models provide confidence that our results are not driven by an over-parameterized CITS model.

Next, we test whether our results are influenced by differential student attrition and mobility. Although NCR schools must continue serving students from their local enrollment area, a few proposed allowing students to enroll in other schools, and students and families may have moved to avoid the reforms. Students moving to other schools may bias our results if those who leave also have systematically higher or lower test scores than those who stay. Some spillover effects could have also occurred if students receive some of the treatment effect from NCR schools, then move into comparison schools. To address these issues, we estimate models that (1) remove all students who transfer between NCR and comparison schools; (2) replace the school fixed effect with a student fixed effect to compare students only with themselves; and (3) remove all student who ever move or leave either NCR or comparison schools. As shown in Appendix Table 16, our results are consistent across these checks.

In addition, it is possible that NCR schools were already testing some reforms on their own in the year before becoming an NCR school (a type of anticipatory effect), or the demoralization from being named a recurring low-performing school may have led to decreased test scores in baseline year before NCR reforms (a type of Ashenfelter dip). In these cases, our results may be influenced by abnormal school performance in the year before NCR reforms. To test for this possibility, we conduct a placebo test by using a CITS model where we recenter *NCRYear* to zero in the year before the baseline year, replace the *AfterNCR* indicator with an "Baseline Year" indicator that switches to one in the baseline year before schools implement NCR reforms, and drop all years after this baseline year. Thus, any significant effects on

*EverNCR*Baseline Year* would suggest that school performance changed in the year before reforms began. Appendix Table 17 shows no significant effect on test scores in the baseline year.

Finally, our results may be biased if either NCR or comparison schools experienced alternative whole-reforms outside of NCR. Using nearly-eligible schools as our comparison group ensures that none of the schools in our sample are implementing any of the other three reform models adopted by North Carolina (transformation, turnaround, or closure). However, while NCR schools were implementing reforms, the state was also piloting a teacher leadership program called Advance Teaching Roles (ATR) in several districts. All schools in these ATR districts would have participated in the pilot including any NCR or comparison schools. To ensure that our results are not driven by ATR reforms, we exclude all ATR schools from our analysis and find that the effect estimates are unchanged (Appendix Table 18). Finally, one district in North Carolina decided to implement NCR in all of its schools in a district-wide intervention called the renewal district. Although the renewal school district had not begun reforms during the period of our study, the district may be distinct in unobserved ways, so we test a sample that excludes all renewal district schools in all years. Excluding the renewal district does not change our results (Appendix Table 18).

Discussion

In this paper, we examined the impact, intervention features, and mediating effects of North Carolina's novel school restart model. Given NCR's tight alignment with ESSA priorities (i.e., increasing local autonomy), our examination provides an important empirical test of the current federal theory of action for school reform. Specifically, we examined whether increased flexibility in the school reform process can improve student outcomes in persistently low-performing schools. Examining NCR also illuminates the theoretical implications of high-

flexibility reforms. Most previous approaches to school reform framed the prescriptive interventions as undesirable, but necessary, consequences and motivated chronically low-performing schools to improve so that they could avoid having to implement required reforms. NCR completely flips this traditional reform logic by marketing the ability to implement reforms as an incentive, requiring chronically low-performing schools to apply for NCR status, and motivating schools to improve in order to retain their restart status. Our examination sheds light on how NCR's reverse approach to reform affects both students and educators.

Using data from the first three years of implementation, we found positive NCR effects in math (0.04 SD), particularly in the second year of implementation (0.11 SD), but no significant effects in ELA, which aligns with a broader literature finding that ELA scores tend to be less sensitive to school-based interventions and more influenced by external factors like language spoken outside the classroom (Charity et al., 2004; Jackson et al., 2014). Our estimates align with findings from a recent meta-analysis of turnaround by Schueler et al. (2021), which found a positive and significant effect of turnaround on math test scores (0.06 SD) and nonsignificant result in ELA. It is also useful to compare the effects we find from NCR to effect estimates from evaluations of previous turnaround efforts in North Carolina. Overall, the mixed results of NCR align with mixed results from evaluations of TALAS (Heissel & Ladd, 2017; Henry & Guthrie, 2019) and NCT (Henry & Harbatkin, 2020), which produced effect sizes ranging from -0.13 SD to 0.09 SD. Though decidedly mixed, these prior evaluations have found evidence that both North Carolina's previous TALAS and NCT turnaround models may have negatively affected student achievement in some years, and we find no such evidence so far under NCR. At the very least, our results suggest that NCR is no less effective than previous turnaround efforts in North Carolina, while also avoiding highly disruptive interventions. These results provide some

evidence that local autonomy may be a useful component of a broader school reform strategy. However, overall, we do not find strong evidence to believe that NCR is more effective at improving student achievement than previous, more prescriptive turnaround models implemented in either North Carolina or nationwide.

Larger and more consistent positive effects after the first year of implementation also aligns with previous literature on school reform, suggesting that longer duration (or greater dosage) is associated with larger effect estimates (Schueler et al., 2021). Our communications with school and district leaders support this dosage effect because administrators frequently mentioned needing time to understand and negotiate what they could and could not do under NCR (Matthews et al., 2022). Instead of a dosage effect, larger estimates in the second year could also be explained by a maturation effect where the NCR policy itself changed over time. However, we found little evidence of a maturation effect. First, the written legislation did not change, and NCDPI leaders did not describe any changes in how they implemented the policy. Second, we found that the effect increased after year one for both the first and second cohort of NCR schools, and if the policy had changed over time, we would have expected instead to see differential effects between the two cohorts. More definitive evidence on dosage versus maturation effects will have to await future research with additional years of data, but positive results within the first three years is an encouraging sign that the NCR model can initially improve student achievement.

In contrast to recent studies that found positive reform effects on non-test-based outcomes (Redding & Nguyen, 2020), we found no evidence of NCR effects on attendance rates, chronic absenteeism, behavior (i.e., reportable offense, long-term suspension), dropout, or graduation. One potential explanation for these null results is that NCR, like many other reform

models, focused on improving student achievement in the short term rather than on non-academic outcomes (Pham et al., 2020). Alternatively, it may take more than 2-3 years for effects on non-test outcomes to be detectable. Nevertheless, the results are not surprising because, under NCR, schools had no direct incentive to improve non-test outcomes. Thus, principals likely focused on academic achievement because school performance ratings in North Carolina are still primarily calculated using test scores. This explanation suggests that even in high-flexibility reform models, policymakers may want to consider incentives for improving non-test outcomes to motivate more holistic approaches to supporting students in low-performing schools.

We also examined NCR school applications to evaluate the relationship between proposed intervention features and student achievement. Schueler and colleagues (2021) found that teacher replacements and extended learning time have positive effects on student achievement, and Fryer (2013) found that charter-school practices (e.g., increased instructional time, high dosage tutoring) raised math scores. We found that one proposed intervention—changing the curriculum—was negatively associated with math performance, suggesting that this sort of large-scale disruption can actually decrease student achievement. Furthermore, unlike Schueler and colleagues (2021), who found that reform models with more intervention features tended to raise student achievement, we found that NCR schools proposing more interventions did not increase student achievement. In fact, we found negative relationships by the third year of implementation. Although our data is limited to proposed interventions and not interventions implemented in practice, this finding implies that schools proposing a larger number of interventions (i.e., the kitchen-sink approach) may be less effective at improving achievement outcomes over time than schools with a more coherent and streamlined approach. These results

imply that future reform planning processes should consider depth over breath and that bold staffing interventions continue to be a high leverage school reform strategy.

Turning to the mediation results, several important pathways help explain the positive NCR effects. We found that 14% of the NCR effect in math can be explained by decreased principal turnover, and 8% can be explained by decreased teacher turnover. These findings help triangulate qualitative evidence from school principals who cite the NCR flexibilities as attractive features of the school for both them and their teaching staff (Matthews et al., 2022). Moreover, this result provides evidence to directly support the NCR theory of action that uses reforms as an incentive rather than relying on sanctions and threats of dismissal to motivate school staff.

Another mediating pathway that supports the NCR theory of action is that 8% of the NCR effect can be explained by alternatively licensed teachers. NCR flexibilities gave school principals autonomy over who they hired, and nearly 60% of NCR school applications proposed flexibility over who could be hired as teachers. Importantly, this result does not imply that alternative certification in general is positively related to student achievement. Rather, we found that alternatively certified teachers *in NCR schools* were associated with higher test scores. This distinction is important because it highlights how giving principals more flexibility may have led to positive outcomes. Multiple principals in our interviews mentioned using NCR flexibilities to hire unlicensed people who they knew would be good teachers and supporting them through alternative certification pathways while they were teaching (Matthews et al., 2022). In these cases, our results support the idea that NCR gave principals enough autonomy to make strategic hiring decisions based on their local knowledge of teaching candidates with high potential. Nevertheless, it is important to note that alternatively licensed teachers have been found to be

more likely to leave teaching (Carver-Thomas & Darling-Hammond, 2019). Thus, future research on NCR schools should explore the longitudinal outcomes of hiring alternatively licensed teachers, with a particular focus on unlicensed teacher retention and its impact on student performance.

Increasing teacher salary is an additional mediating mechanism. The overall increase in NCR teacher salaries appears to be driven primarily by incoming teachers in NCR schools who received higher pay and had higher Praxis scores than teachers hired into comparison schools (Appendix Table 10). Together the results imply that NCR schools used their budgeting flexibilities to invest in teacher pay as a way of recruiting candidates with higher academic qualifications. However, we find no positive effect on the VA score of either entering or exiting teachers. If anything, the effect on VA scores of incoming teachers appears to be negative. If increasing teacher VA scores are not the result of either hiring high VA teachers or dismissing low VA teachers, this would suggest that NCR schools are developing teachers who are already there. Our data do not capture measures to more definitively examine professional development practices in NCR schools, but these results suggest that school and district leaders are making gains in teachers' professional capacity, which prior research has shown to be an essential support for school improvement (Bryk et al., 2010).

Overall, this study finds compelling, but not yet definitive, evidence to support NCR's, and by extension ESSA's, theory of action for school reform. Giving school and district leaders flexibility over reforms in their chronically lowest-performing schools led to positive effects in math, but not in ELA. Because the COVID-19 pandemic limited our current analysis to only the first three years of NCR implementation, we emphasize the need for future analyses with more years of post-pandemic data to better understand long-run effects. Moreover, examining how

NCR schools fared relative to other chronically low-performing schools during the pandemic is an important next step to understanding whether increased flexibility can help school leaders navigate large-scale disruptions. Also, nonsignificant effects on student outcomes outside of test scores suggest that NCR could be improved with incentives to better support students beyond test scores. However, our analysis supports the NCR model as a useful strategy for increasing both teacher and principal retention, and future reform models under ESSA would do well to consider increased flexibility as feature that can be used to retain educators in the schools that need them most.

Endnotes

1. For schools where the NCR start year was not clearly specified in the NCR application, we assumed they began NCR reforms in the year after submitting their application. Results are robust when we exclude the 20 schools with unclear start dates from the analysis (Appendix Table 18).

2. Because student mobility across schools can introduce attrition, we estimate models that exclude all NCR schools proposing school choice and find that our results are robust (Appendix Table 18).

3. The five years of data before NCR began are important to establish credible baseline trends (Bloom, 2003).

4. Note that NCDPI requires students to attend at least 10 instructional days at any time during the school year before they can be included in calculations of chronic absenteeism.

5. In North Carolina, reportable offenses include assault, sexual assault, rape, kidnapping, indecent liberties with a minor, assault involving use of a weapon, possession of a firearm or weapon in violation of the law, possession of a controlled substance in violation of the law (North Carolina General Statute § 115C-288(g)).

6. EverNCR and NCRYear are not included on their own because they are perfectly collinear with the school and year fixed effects.

7. Although the parallel trends assumption is not necessary in the CITS model, coefficient 3 in Equation 1 allows us to directly examine whether pre-treatment trends are significantly different between NCR and comparison schools.

References

Aladjem, D. K., LeFloch, K. C., Zhang, Y., Kurki, A., Boyle, A., Taylor, J. E., Herrmann, S., Uekawa, K., Thomsen, K., & Fashola, O. (2006). Models Matter–The Final Report of the National Longitudinal Evaluation of Comprehensive School Reform. *American Institutes for Research*.

Anrig, G. (2015). *Lessons from school improvement grants that worked*. The Century Foundation. http://www. tcf. org/blog/detail/lessons-from-schoolimprovement-grants-that-worked

Atchison, D. (2020). The impact of priority school designation under ESEA flexibility in New York State. *Journal of Research on Educational Effectiveness*, *13*(1), 121–146.

Baron, R. M., & Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis*, *38*(2), 318–335.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. https://doi.org/10.1093/biomet/93.3.491

Berends, M., Bodilly, S. J., & Kirby, S. N. (2002). *Facing the challenges of whole-school reform: New American Schools after a decade*. Rand Corporation. https://books.google.com/books?hl=en&lr=&id=IAFT_Ln9JwcC&oi=fnd&pg=PR3&dq=

berends+kirby+facing+the+challenges&ots=cy3UELrPXl&sig=-coXvQr4H_wXxWGbvKVG92sCxfU

Bloom, H. S. (2003). Using “Short” Interrupted Time-Series Analysis To Measure The Impacts Of Whole-School Reforms. *Evaluation Review*, *27*(1), 3–49. https://doi.org/10.1177/0193841X02239017

Bonilla, S., & Dee, T. (2017). *The Effects of School Reform Under NCLB Waivers: Evidence from Focus Schools in Kentucky*. National Bureau of Economic Research.

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q., & Luppescu, S. (2010). *Organizing schools for improvement: Lessons from Chicago*. University of Chicago Press.

Callaway, B., & Sant’Anna, P. H. (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

Carver-Thomas, D., & Darling-Hammond, L. (2019). The trouble with teacher turnover: How teacher attrition affects students and schools. *Education Policy Analysis Archives*, *27*(36).

Charity, A. H., Scarborough, H. S., & Griffin, D. M. (2004). Familiarity with school English in African American children and its relation to early reading achievement. *Child Development*, *75*(5), 1340–1356.

Dee, T., & Dizon-Ross, E. (2019). School Performance, Accountability, and Waiver Reforms: Evidence from Louisiana. *Education Evaluation and Policy Analysis*, *41*(3), 316–349.

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

Dixon, L. L., Pham, L. D., Henry, G. T., Corcoran, S. P., & Zimmer, R. (2021). Who Leads Turnaround Schools? Characteristics of Principals in Tennessee’s Achievement School

District and Innovation Zones. *Educational Administration Quarterly*, 0013161X211055702.

Dougherty, S. M., & Weiner, J. M. (2017). The Rhode to Turnaround: The Impact of Waivers to No Child Left Behind on School Performance. *Educational Policy*, 0895904817719520.

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, C., Boyle, A., Upton, R., Tanenbaum, C., & Giffin, J. (2017). School Improvement Grants: Implementation and Effectiveness. NCEE 2017-4013. *National Center for Education Evaluation and Regional Assistance*.

Fryer, R. G. (2013). *Teacher incentives and student achievement: Evidence from New York City public schools*. National Bureau of Economic Research. http://www.nber.org/papers/w16850

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Granados, A. (2017). *Superintendent shares plans for NC's new Achievement School District: WRAL.com*. WRAL-TV. https://www.wral.com/superintendent-shares-plans-for-nc-s-new-achievement-school-district/16807037/

Granados, A., & Hinchcliffe, K. (2018, March 19). Restart program gives some low-performing schools flexibility to help struggling students. *EducationNC*. https://www.ednc.org/restart-program-gives-some-low-performing-schools-flexibility-to-help-struggling-students/

Heissel, J. A., & Ladd, H. F. (2017). School turnaround in North Carolina: A regression discontinuity analysis. *Economics of Education Review*, *62*, 302–320.

Hemelt, S. W., & Jacob, B. A. (2020). How Does an Accountability Program that Targets Achievement Gaps Affect Student Performance? *Education Finance and Policy*, *15*(1), 45–74.

Henry, G. T., & Guthrie, J. E. (2019). The effects of Race to the Top school turnaround in North Carolina. *EdWorkingPapers. Com*.

Henry, G. T., Guthrie, J. E., & Townsend, L. W. (2015). Outcomes and impacts of North Carolina's initiative to turn around the lowest-achieving schools. *The Friday Institute for Educational Innovation, North Carolina State University. Google Scholar*.

Henry, G. T., & Harbatkin, E. (2020). The Next Generation of State Reforms to Improve their Lowest Performing Schools: An Evaluation of North Carolina's School Transformation Intervention. *Journal of Research on Educational Effectiveness*, *13*(4), 702–730.

Henry, G. T., Pham, L. D., Kho, A., & Zimmer, R. (2020). Peeking into the black box of school turnaround: A formal test of mediators and suppressors. *Education Evaluation and Policy Analysis*, *40*(2), 232–256.

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., Redding, S., & Darwin, M. (2008). *Turning Around Chronically Low-Performing Schools. IES Practice Guide*. National Center for Education Evaluation and Regional Assistance. http://eric.ed.gov/?id=ED501241

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics*, *6*(1), 801–825.

Klein, A. (2016). *The Every Student Succeeds Act: An ESSA Overview*. Education Week. https://www.edweek.org/ew/issues/every-student-succeeds-act/index.html?cmp=SOC-SHR-FB

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research, 78*(3), 608–644.

Malen, B., Croninger, R., Muncey, D., & Redmond-Jones, D. (2002). Reconstituting schools:"Testing" the "theory of action." *Educational Evaluation and Policy Analysis, 24*(2), 113–132.

Matthews, G. F., Pham, L. D., Jackson, M., & Singleton, D. T. (2022). *Rebuilding the Box: School Principals Navigating New Reform Environments*. University Council of Educational Administration, Seattle, WA.

NCDPI. (2019). *NCDPI State Test Results (Green Book)—Historical Trends and Results*. https://www.dpi.nc.gov/media/7837/open

NCDPI. (2021). *School Attendance and Student Accounting Manual*. https://www.dpi.nc.gov/media/1258/open

NCDPI. (2016). *Reform Model Information*. https://www.rep.dpi.state.nc.us/app/reform_models/reform.html

NCDPI. (2017). *Every Student Succeeds Act (ESSA) | NC DPI*. https://www.dpi.nc.gov/districts-schools/federal-program-monitoring/every-student-succeeds-act-essa

NCERDC. (2009). *Technical Report #5: Linking Teachers in the Course Membership Data to Teachers in the School Activity Report*. https://childandfamilypolicy.duke.edu/wp-content/uploads/sites/2/2021/11/TECHREPT5.pdf

North Carolina General Statute § 115C-105.37B. https://www.ncleg.net/enactedlegislation/statutes/html/bysection/chapter_115c/gs_115c-105.37b.html

North Carolina General Statute § 115C-288(g), § 115C-288(g). https://www.ncleg.net/enactedlegislation/statutes/html/bysection/chapter_115c/gs_115c-288.html

Pham, L. D. (2022). Why Do We Find These Effects? An Examination of Mediating Pathways Explaining the Effects of School Turnaround. *Journal of Research on Educational Effectiveness, In Press*.

Pham, L. D., Henry, G. T., Kho, A., & Zimmer, R. (2020). Sustainability and maturation of school turnaround: A multi-year evaluation of Tennessee's achievement school district and local innovation zones. *AERA Open*, *6*(2). https://doi.org/10.1177/2332858420922841

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, *66*.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717–731.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891.

Redding, C., & Nguyen, T. D. (2020). The relationship between school turnaround and student outcomes: A meta-analysis. *Educational Evaluation and Policy Analysis*, *42*(4), 493–519.

Schueler, B. E., Asher, C. A., Larned, K. E., Mehrotra, S., & Pollard, C. (2021). Improving Low-Performing Schools: A Meta-Analysis of Impact Evaluation Studies. *American*

Educational Research Journal, 00028312211060855. https://doi.org/10.3102/00028312211060855

Schueler, B. E., & Bleiberg, J. F. (2021). Evaluating Education Governance: Does State Takeover of School Districts Affect Student Achievement? *Journal of Policy Analysis and Management*, pam.22338. https://doi.org/10.1002/pam.22338

Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2013). The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation. *MDRC*.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy*, *10*(4), 535–572.

Strunk, K. O., Marsh, J. A., Hashim, A. K., & Bush-Mecenas, S. (2016). Innovation and a return to the status quo: A mixed-methods study of school reconstitution. *Educational Evaluation and Policy Analysis*, *38*(3), 549–577.

Tyack, D. B., & Cuban, L. (1995). *Tinkering toward utopia a century of public school reform*. Harvard University Press.

U.S. Department of Education. (2009). *Guidance on fiscal yar 2010 school improvement grants under section 1004(3)*. https://www.google.com/webhp?sourceid=chrome-instant&rlz=1C5CHFA_enUS698US698&ion=1&espv=2&ie=UTF-8#q=guidance%20on%20fiscal%20yar%202010%20school%20improvement%20grants%20under%20section%201004(3)

Woessmann, L. (2007). International evidence on school competition, autonomy, and accountability: A review. *Peabody Journal of Education*, *82*(2–3), 473–497.

Zimmer, R., Henry, G. T., & Kho, A. (2017). The Effects of School Turnaround in Tennessee's Achievement School District and Innovation Zones. *Educational Evaluation and Policy Analysis, 39*(4), 670–696.

Table 1. Number of NCR Schools and Districts by Proposed Intervention Feature

	Number of Schools	Number of Districts
Administrator Interventions		
Assistance for School Administrators	34	9
New School Management Organization	8	1
Teacher Interventions		
Flexibility in Teacher Hiring or Assignment	65	19
Additional Resources		
Flexibility in how Existing Funds Are Allocated	34	13
Additional Tutoring	12	8
Wraparound Services	10	5
Other Interventions		
Change Curriculum	49	20
Use Data to Inform Instruction	33	5
Allow Students to Enroll in a Different School	4	1
Common Intervention (Almost Always Proposed Together)		
Human Resource Changes	75	25
Teacher Professional Development	73	21
Extended Learning Time	72	26

Note. Intervention features are coded following definitions from Schueler et al. (2021). Data are from our coding of publicly available improvement plans that schools used to apply for NCR flexibilities. Schools may have proposed multiple implementation features. 111 total NCR school applications are included in our coding. Throughout the analysis, we pool together the three interventions that are almost always proposed together (human resource changes, teacher professional development, extended learning time).

Table 2. Descriptive Characteristics for NCR and Nearly-Eligible Comparison Schools – Before and After Reforms

	Years Before Reforms			Years After Reforms		
	Nearly-Eligible Comparison	NCR		Nearly-Eligible Comparison	NCR	
	Mean (SD)	Mean (SD)	Difference [p-val]	Mean (SD)	Mean (SD)	Difference [p-val]
Student Characteristics						
Female	0.485	0.483	-0.002	0.485	0.479	-0.005
	(0.038)	(0.034)	[0.524]	(0.032)	(0.030)	[0.118]
Asian	0.020	0.019	-0.001	0.023	0.016	-0.007
	(0.041)	(0.026)	[0.765]	(0.055)	(0.022)	[0.158]
Black	0.482	0.534	0.052+	0.472	0.531	0.060+
	(0.242)	(0.229)	[0.082]	(0.235)	(0.234)	[0.057]
Latino/a/x	0.221	0.222	0.001	0.259	0.256	-0.003
	(0.176)	(0.162)	[0.961]	(0.184)	(0.178)	[0.900]
White	0.204	0.182	-0.022	0.175	0.154	-0.021
	(0.189)	(0.167)	[0.330]	(0.175)	(0.148)	[0.315]
Students with Disabilities	0.143	0.151	0.008	0.144	0.150	0.006
	(0.047)	(0.041)	[0.106]	(0.047)	(0.038)	[0.261]
Economically Disadvantaged	0.754	0.756	0.001	0.672	0.664	-0.007
	(0.190)	(0.142)	[0.942]	(0.169)	(0.129)	[0.691]
Multilingual Learner	0.111	0.110	-0.002	0.127	0.132	0.005
	(0.107)	(0.093)	[0.899]	(0.118)	(0.110)	[0.708]
School Characteristics						
Enrollment	517.589	571.353	53.763+	506.086	530.590	24.504
	(234.656)	(242.163)	[0.075]	(242.320)	(237.232)	[0.430]
Rural	0.328	0.277	-0.051	0.314	0.293	-0.021
	(0.470)	(0.448)	[0.368]	(0.465)	(0.456)	[0.730]
Town	0.209	0.113	-0.096*	0.215	0.106	-0.108*
	(0.407)	(0.317)	[0.029]	(0.411)	(0.309)	[0.019]
Suburb	0.086	0.178	0.092*	0.091	0.176	0.084+
	(0.280)	(0.383)	[0.027]	(0.288)	(0.381)	[0.060]
City	0.377	0.432	0.054	0.380	0.426	0.045
	(0.485)	(0.496)	[0.386]	(0.486)	(0.496)	[0.484]
Elementary School	0.689	0.703	0.014	0.694	0.697	0.003
	(0.463)	(0.457)	[0.809]	(0.461)	(0.461)	[0.961]
Middle School	0.266	0.218	-0.048	0.262	0.213	-0.049
	(0.442)	(0.413)	[0.382]	(0.440)	(0.410)	[0.375]
High School	0.030	0.077	0.047	0.030	0.090	0.061+
	(0.171)	(0.268)	[0.104]	(0.170)	(0.288)	[0.068]
N Schools	135	111		135	111	

Note. The comparison group is all schools nearly-eligible for NCR. For comparison schools, the years before NCR are 2011-12 through 2015-16 and the years after are 2016-17 through 2018-19. For NCR schools, the years before are all years before the school implements reforms and after includes all years after NCR reforms began. Significance stars are from descriptive t-tests comparing NCR and comparison schools.

Table 3. CITS Effect Estimates on Math and ELA Test Scores using Different Comparison Groups

Outcome:	Math Test Scores				ELA Test Scores			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Comparison Group:	Nearly Eligible	All Eligible	Rescinded	Future	Nearly Eligible	All Eligible	Rescinded	Future
Ever NCR*NCR Year	-0.009	-0.001	0.001	-0.008	-0.000	0.001	0.001	-0.001
	(0.005)	(0.003)	(0.003)	(0.005)	(0.002)	(0.002)	(0.002)	(0.003)
Year 1	-0.017+	-0.010	-0.001	-0.031+	-0.006	-0.010	-0.004	-0.020
	(0.010)	(0.012)	(0.012)	(0.017)	(0.008)	(0.010)	(0.010)	(0.013)
Year 2	-0.006	0.010	0.012	-0.034	0.009	0.007	0.007	-0.011
	(0.013)	(0.015)	(0.016)	(0.023)	(0.011)	(0.013)	(0.013)	(0.015)
Year 3	0.083+	0.103*	0.088*	0.004	-0.002	0.005	0.007	-0.013
	(0.042)	(0.043)	(0.044)	(0.053)	(0.015)	(0.017)	(0.017)	(0.020)
Ever NCR*Year 1	0.024+	0.017	0.008	0.035+	0.012	0.017	0.011	0.027+
	(0.012)	(0.014)	(0.014)	(0.018)	(0.010)	(0.012)	(0.012)	(0.014)
Ever NCR*Year 2	0.113**	0.097*	0.095*	0.139**	0.002	0.005	0.004	0.022
	(0.040)	(0.041)	(0.041)	(0.049)	(0.012)	(0.013)	(0.013)	(0.016)
Ever NCR*Year 3	0.108	0.088	0.103	0.201*	-0.013	-0.019	-0.022	-0.002
	(0.095)	(0.096)	(0.096)	(0.100)	(0.017)	(0.019)	(0.019)	(0.021)
N (Student-Year)	492553	412662	401236	216544	480134	395204	383965	292083
R-Squared	0.788	0.779	0.783	0.778	0.888	0.886	0.890	0.887
Student Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School and Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note. Standard errors in parentheses. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. Nearly eligible schools serve as our preferred comparison group. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4. CITS Effect Estimates on Other Student Outcomes

	(1) Math	(2) ELA	(3) Attendance Rate	(4) Chronic Absenteeism	(5) Report-able Offense	(6) Long Term Suspension	(7) Dropout	(8) Graduate
Ever NCR*NCR Year	-0.009	-0.000	0.000	-0.000	-0.000	0.000	-0.000	-0.001
	(0.005)	(0.002)	(0.001)	(0.003)	(0.003)	(0.001)	(0.000)	(0.000)
Year 1	-0.017+	-0.006	-0.000	-0.002	0.007	0.004	-0.001	0.002*
	(0.010)	(0.008)	(0.002)	(0.009)	(0.007)	(0.003)	(0.001)	(0.001)
Year 2	-0.006	0.009	0.002	-0.008	0.013	0.005	-0.001	-0.003
	(0.013)	(0.011)	(0.003)	(0.011)	(0.011)	(0.004)	(0.001)	(0.002)
Year 3	0.083+	-0.002	0.004	-0.027	-0.004	0.005	-0.001	-0.004
	(0.042)	(0.015)	(0.004)	(0.018)	(0.015)	(0.005)	(0.001)	(0.004)
Ever NCR*Year 1	0.024+	0.012	0.002	-0.013	-0.004	-0.003	0.001	-0.002
	(0.012)	(0.010)	(0.003)	(0.013)	(0.009)	(0.002)	(0.001)	(0.002)
Ever NCR*Year 2	0.113**	0.002	0.002	-0.019	-0.013	-0.005+	0.000	0.004
	(0.040)	(0.012)	(0.003)	(0.013)	(0.014)	(0.003)	(0.002)	(0.004)
Ever NCR*Year 3	0.108	-0.013	0.000	-0.013	-0.006	-0.003	-0.003	0.011
	(0.095)	(0.017)	(0.004)	(0.016)	(0.017)	(0.003)	(0.003)	(0.008)
N (Student-Year)	492553	480134	397682	397682	169880	169880	703328	703328
R-Squared	0.788	0.888	0.361	0.216	0.893	0.008	0.017	0.197
Student Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School and Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note. Standard errors in parentheses. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, and ML status. For test score outcomes, covariates also include prior year lagged test score. Models with dichotomous outcomes are interpreted as linear probability models. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5. CITS Estimates by Proposed Intervention Feature

	(1) Math	(2) ELA
New Manager*After NCR	-0.039	0.017
	(0.033)	(0.014)
Administrator Assistance*After NCR	0.008	-0.010
	(0.036)	(0.013)
Replace Teachers*After NCR	0.014	-0.001
	(0.031)	(0.016)
New Funding*After NCR	-0.005	-0.004
	(0.025)	(0.011)
Additional Tutoring*After NCR	-0.022	-0.002
	(0.048)	(0.015)
Wraparound Services*After NCR	-0.018	-0.014
	(0.037)	(0.012)
Change Curriculum*After NCR	-0.058*	0.012
	(0.027)	(0.012)
Use Data*After NCR	-0.047+	-0.002
	(0.026)	(0.010)
School Choice*After NCR	-0.076+	-0.048
	(0.044)	(0.048)
Common Intervention*After NCR	0.054	0.001
	(0.035)	(0.018)
N (Student-Year)	479427	467616
R-Squared	0.790	0.889
Student Covariates	Yes	Yes
School and Year Fixed Effect	Yes	Yes

Note. Standard errors in parentheses. Only the interaction between each intervention feature and the after NCR indicator are reported for clarity, but results are from estimating the full CITS model. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. The common intervention is an indicator for whether schools proposed either human resource changes, teacher professional development, or extended learning time. These three intervention features had to be pooled together because they were almost always proposed together. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1. Schema Illustrating Statistical Mediational Analysis

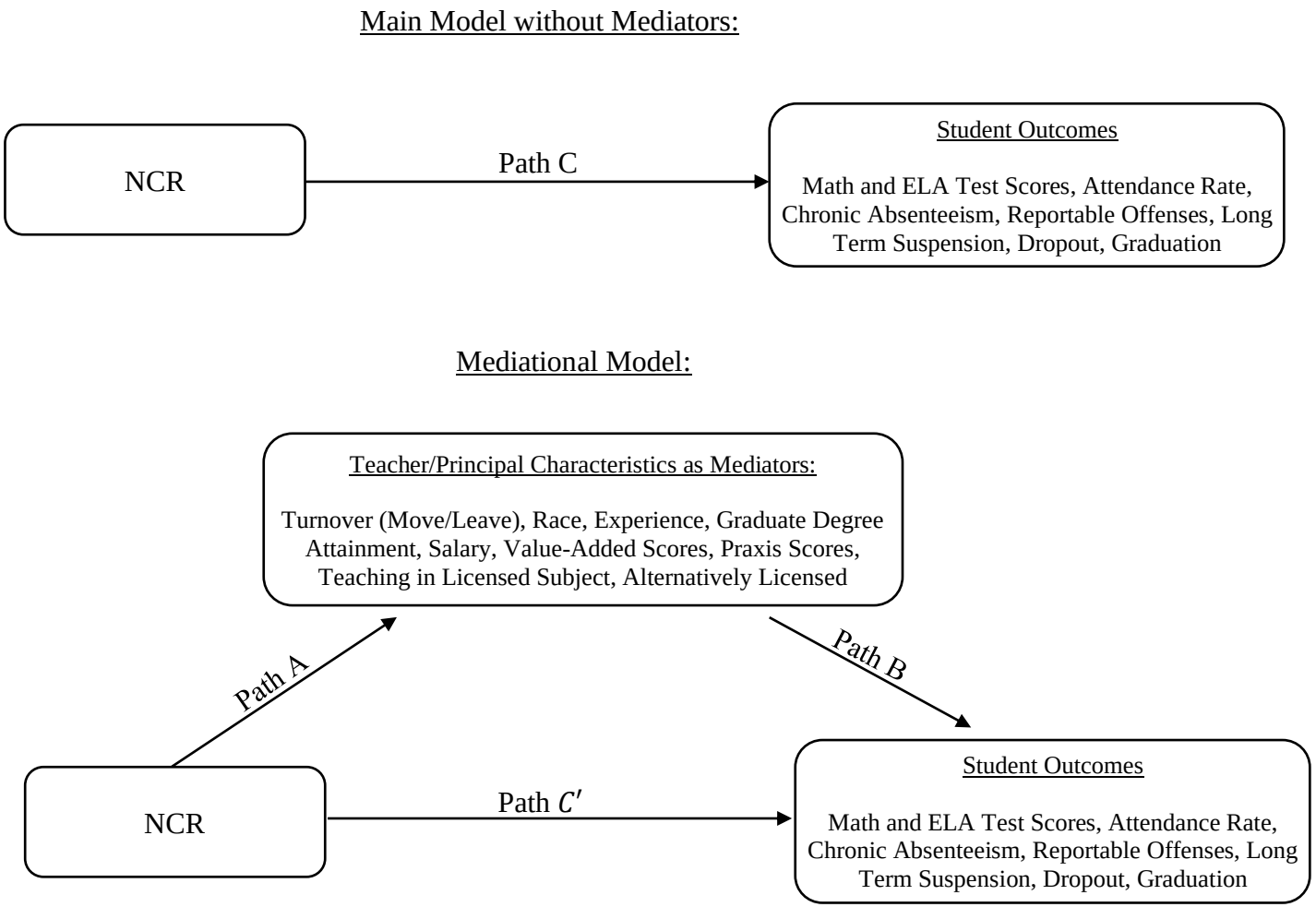
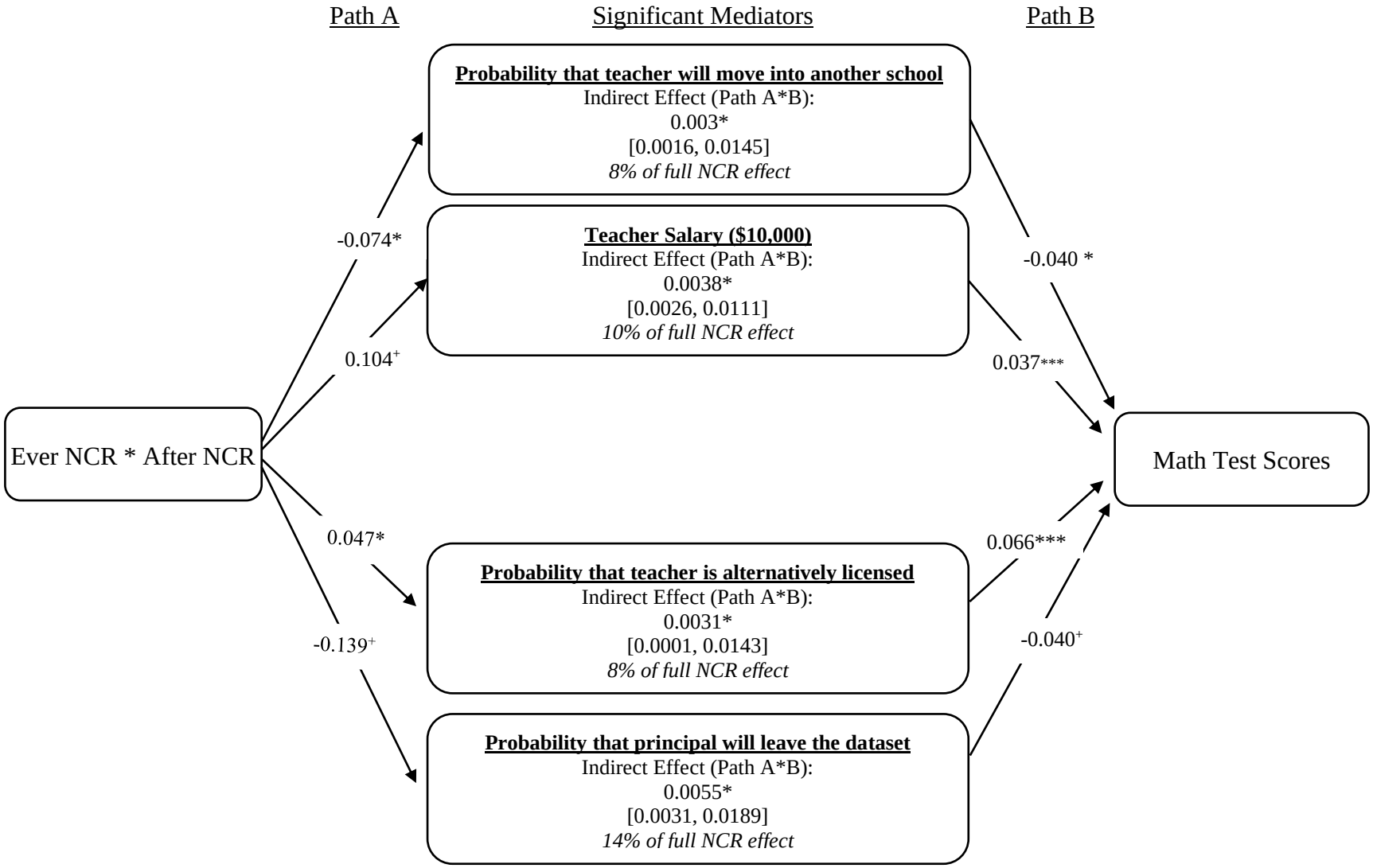


Figure 2. Depiction of Statistically Significant Mediating Mechanisms that Explain the Effect of NCR Reforms on Math Test Scores



Note. For clarity, this figure depicts the NCR effect using only the interaction between the EverNCR and AfterNCR indicators, but these results are from estimating the full CITS model with student level controls and school and year fixed effects. See Equation 1. 95% bias-correct confidence intervals from bootstrapping are shown in brackets. Percent of full NCR effect is calculated using a 0.04 SD effect size. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Tables

Appendix Table 1. Descriptive Characteristics for Each NCR Cohort – Before and After Reforms

	Years Before Reforms			Years After Reforms		
	(1)	(2)	(3)	(4)	(5)	(6)
	Cohort 1	Cohort 2	Cohort 3	Cohort 1	Cohort 2	Cohort 3
Student Demographics						
Female	0.481	0.479	0.481	0.488	0.480	0.472
	(0.031)	(0.033)	(0.036)	(0.024)	(0.030)	(0.031)
Asian	0.009	0.015	0.026	0.011	0.014	0.024
	(0.014)	(0.022)	(0.030)	(0.016)	(0.020)	(0.030)
Black	0.529	0.540	0.520	0.536	0.538	0.506
	(0.308)	(0.230)	(0.216)	(0.309)	(0.227)	(0.222)
Latino/a/x	0.367	0.201	0.265	0.373	0.230	0.290
	(0.265)	(0.140)	(0.175)	(0.265)	(0.153)	(0.187)
White	0.062	0.199	0.148	0.050	0.171	0.142
	(0.070)	(0.172)	(0.142)	(0.053)	(0.154)	(0.139)
Students with Disabilities	0.158	0.144	0.158	0.142	0.147	0.164
	(0.059)	(0.040)	(0.040)	(0.031)	(0.039)	(0.037)
Economically Disadvantaged	0.835	0.724	0.740	0.822	0.638	0.681
	(0.101)	(0.136)	(0.137)	(0.092)	(0.121)	(0.114)
Multilingual Learner	0.183	0.085	0.133	0.194	0.112	0.172
	(0.134)	(0.069)	(0.100)	(0.126)	(0.095)	(0.132)
School Characteristics						
Enrollment	625.389	585.974	536.542	572.056	531.227	508.737
	(194.782)	(251.382)	(232.237)	(174.526)	(247.574)	(228.259)
Rural	0.333	0.313	0.184	0.333	0.318	0.184
	(0.485)	(0.465)	(0.389)	(0.485)	(0.468)	(0.393)
Town	0.000	0.119	0.105	0.000	0.121	0.105
	(0.000)	(0.325)	(0.308)	(0.000)	(0.328)	(0.311)
Suburb	0.000	0.194	0.200	0.000	0.197	0.184
	(0.000)	(0.396)	(0.401)	(0.000)	(0.399)	(0.393)
City	0.667	0.373	0.511	0.667	0.364	0.526
	(0.485)	(0.485)	(0.501)	(0.485)	(0.483)	(0.506)
Elementary School	0.833	0.672	0.737	0.833	0.667	0.737
	(0.383)	(0.470)	(0.442)	(0.383)	(0.473)	(0.446)
Middle School	0.000	0.239	0.211	0.000	0.242	0.211
	(0.000)	(0.427)	(0.409)	(0.000)	(0.430)	(0.413)
High School	0.167	0.090	0.053	0.167	0.091	0.053
	(0.383)	(0.286)	(0.224)	(0.383)	(0.289)	(0.226)
N Schools	6	67	38	6	67	38

Note. Standard deviations in parentheses. For NCR schools, the years before are all years before the school implements reforms and after includes all years after NCR reforms began.

Appendix Table 2. Definitions of Interventions Features from Schueler at al. (2021)

Intervention Feature	Definition
New Funding	there was a documented source of additional funding
Governance Change	governance of schools was transferred from the traditional locally elected school board to another party, such as a state takeover of a district
Change in School Manager	day-to-day management of treated units was transferred to a new group, such as a charter management organization
Human Resource Change	there were changes to how teachers were managed, paid, or evaluated, including flexibility from collective bargaining agreements
Teacher Professional Development	teachers were provided with professional development
Administrator Technical Assistance	school or district administrators received supports in the form of professional development or technical assistance
Teacher Replacements	at least 35% of teachers were replaced or the study authors otherwise called out teacher replacements as a major part of the intervention
Principal Replacements	at least 50% of school principals were replaced or leadership replacement was explicitly described as a major part of the intervention
Extended Learning Time	students received additional hours or days of instruction
Tutoring	some students in treated schools received small group or individualized tutoring
Curricular Change	schools changed their curricula as part of the intervention
Data Use	school staff used student data to inform instruction
Wraparound Services	schools or districts provided noninstructional services to students and families, such as counseling, health services, or food support
School Choice	parents could send select schools other than that assigned by neighborhood

Note. Definitions are reproduced from Schueler et al. (2021) on page 10.

Appendix Table 3. Power Analysis by Cohort of NCR Schools

	Number of NCR Schools	Number of Comparison Schools	Power
Analyses that pool all NCR schools	111	135	0.877
Analyses using only cohort 1 NCR schools	6	135	0.159
Analyses using only cohort 2 NCR schools	67	135	0.763
Analyses using only cohort 3 NCR schools	38	135	0.586

Note. Clustered design with students clustered within schools. Two-tailed, $\alpha = .05$. We specified cluster sizes of 682 students per cluster/school based on the average enrollment of NCR schools in our sample.

Appendix Table 4. CITS Effect Estimates on Student Outcomes with False Discovery Rate Adjusted *p*-values

	(1) Math	(2) ELA	(3) Attendance Rate	(4) Chronic Absenteeism	(5) Report-able Offense	(6) Long Term Suspension	(7) Dropout	(8) Graduate
Ever NCR*Year 1	0.024+	0.012	0.002	-0.013	-0.004	-0.003	0.001	-0.002
Unadjusted p-value	(0.056)	(0.232)	(0.552)	(0.309)	(0.664)	(0.152)	(0.343)	(0.285)
FDR-adjusted p-value	[0.215]	[0.215]	[1.000]	[0.215]	[1.000]	[0.215]	[0.215]	[0.215]
Ever NCR*Year 2	0.113**	0.002	0.002	-0.019	-0.013	-0.005+	0.000	0.004
Unadjusted p-value	(0.005)	(0.879)	(0.512)	(0.140)	(0.353)	(0.089)	(0.812)	(0.379)
FDR-adjusted p-value	[0.040]	[1.000]	[0.249]	[0.215]	[0.215]	[0.215]	[1.000]	[0.216]
Ever NCR*Year 3	0.108	-0.013	0.000	-0.013	-0.006	-0.003	-0.003	0.011
Unadjusted p-value	(0.258)	(0.450)	(0.917)	(0.415)	(0.718)	(0.287)	(0.254)	(0.192)
FDR-adjusted p-value	[0.215]	[0.228]	[1.000]	[0.222]	[1.000]	[0.215]	[0.215]	[0.215]
N (Student-Year)	492553	480134	397682	397682	169880	169880	703328	703328
R-Squared	0.788	0.888	0.361	0.216	0.893	0.008	0.017	0.197
Student Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School and Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note. Standard errors in parentheses. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, and ML status. For test score outcomes, covariates also include prior year lagged test score. Models with dichotomous outcomes are interpreted as linear probability models. FDR is false discovery rate (Benjamini et al., 2006). + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 5. CITS Heterogeneous Effects: Outcome is Math Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Moderator:	Female	Black	Latino/a/x	SWD	EDS	EL	Elementary School	City
Ever NCR*NCR Year	-0.009***	-0.010+	-0.010***	-0.002	-0.017+	-0.010**	-0.012*	-0.011**
	(0.001)	(0.005)	(0.001)	(0.007)	(0.009)	(0.002)	(0.004)	(0.002)
Ever NCR*After NCR	0.016	0.030	0.005	-0.012	0.042	0.013	0.007	0.025*
	(0.013)	(0.021)	(0.018)	(0.030)	(0.027)	(0.013)	(0.023)	(0.010)
Moderator	-0.017	-0.016	-0.005	0.079	-0.021	-0.025	-0.182	-0.025
	(0.013)	(0.026)	(0.015)	(0.164)	(0.018)	(0.033)	(0.127)	(0.028)
Moderator*NCR Year	-0.006	-0.006	-0.008+	0.082	-0.016	-0.015	-0.008	-0.006
	(0.005)	(0.013)	(0.004)	(0.057)	(0.017)	(0.013)	(0.005)	(0.006)
Moderator*Ever NCR	0.007	0.009	0.004	-0.034	0.014	0.037+	0.128	0.001
	(0.008)	(0.021)	(0.016)	(0.091)	(0.015)	(0.019)	(0.107)	(0.041)
Moderator*After NCR	0.037	-0.030	0.028	-0.401	-0.015	-0.037	-0.012	-0.005
	(0.025)	(0.050)	(0.021)	(0.284)	(0.055)	(0.054)	(0.019)	(0.024)
Moderator*Ever NCR*NCR Year	0.002	0.004	0.008+	-0.051	0.011	0.016	0.008	0.008
	(0.003)	(0.010)	(0.004)	(0.036)	(0.012)	(0.009)	(0.006)	(0.005)
Moderator*Ever NCR*After NCR	-0.017	-0.049	0.001	0.156	-0.048	-0.068	-0.002	-0.043
	(0.015)	(0.051)	(0.028)	(0.164)	(0.046)	(0.037)	(0.027)	(0.026)
N (Student-Year)	492553	492553	492553	492553	492553	492553	492553	492553
R-Squared	0.787	0.788	0.787	0.788	0.788	0.788	0.787	0.787
Student Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School and Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note. Standard errors in parentheses. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 6. CITS Heterogeneous Effects: Outcome is ELA Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Moderator:	Female	Black	Latino/a/x	SWD	EDS	EL	Elementary School	City
Ever NCR*NCR Year	0.000	0.002	-0.001	0.004	0.003	-0.000	-0.000	-0.000
	(0.002)	(0.004)	(0.002)	(0.006)	(0.004)	(0.003)	(0.004)	(0.003)
Ever NCR*After NCR	0.005	-0.003	0.011	-0.012	-0.010	0.007	0.007	0.008
	(0.013)	(0.021)	(0.013)	(0.030)	(0.022)	(0.016)	(0.022)	(0.019)
Moderator	-0.002	0.011	-0.002	0.098	0.004	0.025+	0.009	-0.039**
	(0.010)	(0.008)	(0.009)	(0.116)	(0.006)	(0.012)	(0.018)	(0.010)
Moderator*NCR Year	-0.001	0.005	-0.003	0.047	0.003	0.002	-0.002	-0.001
	(0.004)	(0.003)	(0.003)	(0.039)	(0.003)	(0.004)	(0.002)	(0.001)
Moderator*Ever NCR	-0.002	-0.010	0.009	-0.031	-0.010	0.020	-0.038	0.021
	(0.006)	(0.009)	(0.011)	(0.064)	(0.012)	(0.026)	(0.022)	(0.022)
Moderator*After NCR	0.019	-0.025	0.018	-0.272	-0.020	-0.026	0.008	0.005
	(0.019)	(0.014)	(0.013)	(0.192)	(0.012)	(0.022)	(0.010)	(0.004)
Moderator*Ever NCR*NCR Year	-0.001	-0.005	0.004	-0.032	-0.005	0.002	0.001	0.000
	(0.002)	(0.003)	(0.004)	(0.024)	(0.003)	(0.006)	(0.003)	(0.003)
Moderator*Ever NCR*After NCR	0.001	0.018	-0.023	0.133	0.021	-0.021	-0.005	-0.008
	(0.011)	(0.013)	(0.019)	(0.106)	(0.017)	(0.035)	(0.015)	(0.015)
N (Student-Year)	480134	480134	480134	480134	480134	480134	480134	480134
R-Squared	0.888	0.888	0.888	0.888	0.888	0.888	0.888	0.888
Student Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School and Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note. Standard errors in parentheses. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 7. CITS Results by Cohort

	Math			ELA		
	(1)	(2)	(3)	(4)	(5)	(6)
	Cohort 1	Cohort 2	Cohort 3	Cohort 1	Cohort 2	Cohort 3
Ever NCR*NCR Year (Centered)	0.003	-0.015	-0.023*	0.002	-0.001	-0.002
	(0.018)	(0.011)	(0.011)	(0.010)	(0.002)	(0.002)
Ever NCR*Year 1	0.023	-0.012	-0.001	-0.006	-0.001	0.017
	(0.040)	(0.039)	(0.110)	(0.025)	(0.007)	(0.009)
Ever NCR*Year 2	0.040	0.071		-0.023	0.016+	
	(0.055)	(0.076)		(0.037)	(0.010)	
Ever NCR*Year 3	0.047			-0.035		
	(0.164)			(0.049)		
N (Student-Year)	173655	359570	281850	172497	262561	207275
R-Squared	0.807	0.566	0.557	0.900	0.900	0.904
Student Covariates	Yes	Yes	Yes	Yes	Yes	Yes
School and Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes

Note. Standard errors in parentheses. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. The Year 1; Year 2; and Year 3 indicators are omitted because they are perfectly collinear with the year fixed effect when examining each cohort separately. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 8. CITS Estimates by Number of Intervention Features Proposed in the NCR Improvement Plan

	(1) Math	(2) ELA
Sum of Proposed Interventions*NCR Year (Centered)	-0.001	-0.000
	(0.001)	(0.000)
Year 1	-0.000	-0.000
	(0.007)	(0.006)
Year 1 * Sum of Proposed Interventions	-0.002	-0.003
	(0.004)	(0.002)
Year 2	0.030*	0.002
	(0.015)	(0.008)
Year 2 * Sum of Proposed Interventions	0.002	-0.004
	(0.007)	(0.003)
Year 3	0.076+	-0.022+
	(0.039)	(0.012)
Year 3 * Sum of Proposed Interventions	-0.025*	-0.002
	(0.011)	(0.004)
N (Student-Year)	492553	480134
R-Squared	0.788	0.888
Student Covariates	Yes	Yes
School and Year Fixed Effect	Yes	Yes

Note. Standard errors in parentheses. The sum of proposed interventions and year trend main effect are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 9. Mediating Effects of Teacher and Principal Characteristics in NCR Schools on Math Test Scores

	Path A	Path B	Indirect Effect
Teachers Characteristics as Mediators			
Will Move	-0.074*	-0.040***	0.003*
	(0.030)	(0.012)	[0.0016, 0.0145]
Will Leave	-0.033+	-0.025*	0.0008
	(0.019)	(0.012)	[-0.0049, 0.0112]
Nonwhite	0.049**	-0.023+	-0.0011
	(0.019)	(0.014)	[-0.0052, -0.0006]
Years of Experience	0.002	0.002***	<0.0001
	(0.366)	(0.001)	[-0.0004, 0.0034]
Graduate Degree	-0.023	0.019+	-0.0004
	(0.022)	(0.010)	[-0.0046, 0.0001]
Salary ($10,000)	0.104+	0.037***	0.0038*
	(0.054)	(0.006)	[0.0026, 0.0111]
Value Added Score	-0.026***	-0.257	0.0068+
	(0.006)	(0.169)	[-0.0144, 0.0009]
Praxis Score	0.021	0.034***	0.0007
	(0.035)	(0.007)	[-0.0022, 0.0082]
Teaching in Licensed Subject	0.092***	-0.033*	-0.0030+
	(0.019)	(0.014)	[-0.0107, -0.0014]
Alternatively Licensed	0.047*	0.066***	0.0031*
	(0.019)	(0.014)	[0.0001, 0.0143]
Principal Characteristics as Mediators			
Will Move	-0.012	-0.002	0.0000
	(0.091)	(0.021)	[-0.0053, 0.0062]
Will Leave	-0.139+	-0.040+	0.0055*
	(0.079)	(0.023)	[0.0031, 0.0189]
Nonwhite	-0.046	-0.066**	0.0031
	(0.068)	(0.024)	[-0.0018, 0.0210]
Years of Experience	0.863	0.004	0.0035
	(0.889)	(0.003)	[-0.0008, 0.0254]
Doctorate Degree	0.012	-0.001	0.0000
	(0.060)	(0.051)	[-0.0118, 0.0100]
Salary ($10,000)	0.295	0.032***	0.0093
	(0.231)	(0.009)	[-0.0119, 0.0247]
Administrator License	-0.062	0.036	-0.0022
	(0.064)	(0.034)	[-0.0183, 0.0012]

Note. Standard errors in parentheses. All standard errors are clustered at the school level. Mediational models include all CITS variables and covariates from Equation 1: indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. 95% bias-corrected confidence intervals from our bootstrap procedure are shown in brackets. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 10. Mediating Effects of Teacher and Principal Characteristics in NCR Schools on Math Test Scores – Incoming Educators Only

	Path A	Path B	Indirect Effect
Characteristics of Incoming Teachers			
Will Move	-0.018	-0.031+	0.0006
	(0.048)	(0.016)	[-0.0024, 0.0105]
Will Leave	-0.026	-0.025	0.0006
	(0.044)	(0.019)	[-0.0025, 0.0113]
Nonwhite	-0.059	-0.001	0.0001
	(0.046)	(0.016)	[-0.0045, 0.0179]
Years of Experience	0.601	0.003***	0.0019
	(0.859)	(0.001)	[-0.0006, 0.0197]
Graduate Degree	-0.027	0.045***	-0.0012
	(0.048)	(0.013)	[-0.0139, 0.0007]
Salary ($10,000)	0.405**	0.033***	0.0135*
	-0.135	-0.005	[0.0056, 0.0373]
Value Added Score	-0.044***	0.897***	-0.0394***
	(0.010)	(0.073)	[-0.0413, -0.0413]
Praxis Score	0.115	0.029**	0.0034*
	(0.094)	(0.009)	[0.0009, 0.0152]
Teaching in Licensed Subject	0.013	-0.017	-0.0002
	(0.036)	(0.019)	[-0.0096, 0.0053]
Alternatively Licensed	0.003	0.047**	0.0002
	(0.051)	(0.015)	[-0.0054, 0.0194]
Characteristics of Incoming Principals			
Will Move	0.123	0.026	0.0031
	(0.115)	(0.045)	[-0.3397, 0.1176]
Will Leave	0.096	-0.070	-0.0067
	(0.110)	(0.048)	[-0.0992, 0.3892]
Nonwhite	-0.305	-0.088*	0.0269
	(0.227)	(0.038)	[-0.0048, 0.4032]
Years of Experience	1.458	0.004	0.0063
	(2.604)	(0.004)	[-0.1855, 0.1255]
Doctorate Degree	-0.239	-0.018	0.0043
	(0.167)	(0.051)	[-0.1311, 0.1395]
Salary ($10,000)	1.172+	0.020	0.0233
	(0.666)	(0.013)	[-0.3699, 0.1526]
Administrator License	-0.369+	0.015	-0.0056
	(0.194)	(0.047)	[-0.2455, 0.0593]

Note. Standard errors in parentheses. All standard errors are clustered at the school level. Mediational models include all CITS variables and covariates from Equation 1: indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. 95% bias-corrected confidence intervals from our bootstrap procedure are shown in brackets. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 11. Mediating Effects of Teacher and principal Characteristics in NCR Schools on Math Test Scores – Outgoing Educators Only

	Path A	Path B	Indirect Effect
Characteristics of Outgoing Teachers			
Nonwhite	0.014	0.005	0.0001
	(0.059)	(0.016)	[-0.0016, 0.0029]
Years of Experience	-1.500+	0.002*	-0.0024
	(0.906)	(0.001)	[-0.0079, 0.0002]
Graduate Degree	-0.023	0.012	-0.0003
	(0.055)	(0.013)	[-0.0041, 0.0009]
Salary ($10,000)	-0.027	0.030***	-0.0008
	-0.141	-0.006	[-0.0085, 0.0063]
Value Added Score	0.013*	-0.341	-0.0044
	(0.006)	(0.223)	[-0.0135, 0.0010]
Praxis Score	-0.076	0.022*	-0.0017
	(0.119)	(0.009)	[-0.0083, 0.0024]
Teaching in Licensed Subject	0.092*	-0.013	-0.0012
	(0.040)	(0.015)	[-0.0041, 0.0016]
Alternatively Licensed	-0.031	0.095***	-0.0029
	(0.044)	(0.016)	[-0.0114, 0.0051]
Characteristics of Outgoing Principals			
Nonwhite	0.151	-0.076*	-0.0115
	(0.201)	(0.030)	[-0.0472, 0.0155]
Years of Experience	1.157	0.008*	0.0088
	(2.108)	(0.003)	[-0.0154, 0.0515]
Doctorate Degree	-0.018	-0.009	0.0002
	(0.161)	(0.050)	[-0.0140, 0.0231]
Salary ($10,000)	0.332	0.021	0.0069
	(0.884)	(0.013)	[-0.0240, 0.0883]
Administrator License	-0.063	0.002	-0.0001
	(0.164)	(0.044)	[-0.0162, 0.0107]

Note. Standard errors in parentheses. All standard errors are clustered at the school level. Mediational models include all CITS variables and covariates from Equation 1: indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. 95% bias-corrected confidence intervals from our bootstrap procedure are shown in brackets. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 12. Alternative CITS Specifications

	Preferred Model		No Covariates		Replace School FE with School Covariates		Replace School FE with District FE		Weight by Enrollment	
	(1) Math	(2) ELA	(3) Math	(4) ELA	(5) Math	(6) ELA	(7) Math	(8) ELA	(9) Math	(10) ELA
Ever NCR*NCR Year (Centered)	-0.009	-0.000	-0.020+	-0.000	-0.022+	0.000	-0.021	-0.000	-0.007*	0.001
	(0.005)	(0.002)	(0.012)	(0.002)	(0.012)	(0.002)	(0.013)	(0.002)	(0.003)	(0.002)
Year 1	-0.017+	-0.006	-0.045*	-0.006	-0.049*	-0.005	-0.047*	-0.006	-0.010	0.002
	(0.010)	(0.008)	(0.021)	(0.008)	(0.021)	(0.008)	(0.021)	(0.011)	(0.011)	(0.008)
Year 2	-0.006	0.009	-0.041	0.009	-0.049	0.012	-0.047	0.010	0.002	0.018
	(0.013)	(0.011)	(0.043)	(0.011)	(0.043)	(0.011)	(0.040)	(0.010)	(0.014)	(0.012)
Year 3	0.083+	-0.002	-0.027	-0.002	-0.031	0.002	-0.030	-0.001	0.102+	0.006
	(0.042)	(0.015)	(0.099)	(0.015)	(0.099)	(0.015)	(0.114)	(0.014)	(0.057)	(0.017)
Ever NCR*Year 1	0.024+	0.012	0.033	0.012	0.034	0.012	0.032	0.013	0.019	0.005
	(0.012)	(0.010)	(0.035)	(0.010)	(0.035)	(0.010)	(0.029)	(0.014)	(0.013)	(0.011)
Ever NCR*Year 2	0.113**	0.002	0.107*	0.003	0.115*	0.001	0.111	0.003	0.109+	-0.005
	(0.040)	(0.012)	(0.051)	(0.012)	(0.052)	(0.012)	(0.110)	(0.013)	(0.057)	(0.013)
Ever NCR*Year 3	0.108	-0.013	0.369**	-0.012	0.310**	-0.010	0.341+	-0.014	0.070	-0.025
	(0.095)	(0.017)	(0.116)	(0.017)	(0.109)	(0.019)	(0.171)	(0.013)	(0.099)	(0.017)
N (Student-Year)	492553	480134	472886	481148	469330	477580	469330	477580	492553	480134
R-Squared	0.788	0.888	0.557	0.888	0.557	0.888	0.559	0.888	0.779	0.887
Student Covariates	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note. Standard errors in parentheses. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. FE stands for fixed effect. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 13. Unadjusted Mean Differences in Math and ELA Test Scores between NCR and Comparison Schools

	(1) Math	(2) ELA
NCR	0.02**	0.01
	(0.01)	(0.01)
N	111957	190033

Note. . + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ Standard errors in parentheses, clustered at the school level. Models include no covariates.

Appendix Table 14. Difference-in-Differences Models with and without Covariates

	(1) Math	(2) ELA	(3) Math	(4) ELA
Ever NCR	-0.02***	-0.02	-0.00	-0.01
	(0.00)	(0.02)	(0.00)	(0.01)
After NCR	0.00	-0.01	0.00	-0.01
	(0.00)	(0.01)	(0.00)	(0.01)
Ever NCR * After NCR	0.04***	0.03	0.02**	0.02
	(0.01)	(0.02)	(0.01)	(0.01)
Covariates	No	No	Yes	Yes
N	553397	631095	552030	551494

Note. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ Standard errors in parentheses, clustered at the school level. Covariates include indicators for gender, race, SWD, ED, and ML status. Models include no fixed effects.

Appendix Table 15. CITS Models that Do Not Include School or Year Fixed Effects and Do Not Use Multiple Indicators for Each Year After NCR Began

	(1) Math	(2) ELA	(3) Math	(4) ELA
Ever NCR	-0.02	-0.02	-0.01	-0.00
	(0.02)	(0.02)	(0.02)	(0.02)
NCR Year	0.01	0.00	0.01	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Ever NCR * NCR Year	-0.00	0.00	-0.00	0.00
	(0.01)	(0.00)	(0.00)	(0.00)
After NCR	-0.02	-0.02	-0.02	-0.02*
	(0.01)	(0.01)	(0.01)	(0.01)
Ever NCR * After NCR	0.05**	0.02	0.04*	0.02
	(0.02)	(0.02)	(0.02)	(0.01)
Covariates	No	No	Yes	Yes
N	553397	631095	552030	629508

Note. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ Standard errors in parentheses, clustered at the school level. Covariates include indicators for gender, race, SWD, ED, and ML status.

Appendix Table 16. Checks for Student Mobility and Attrition

	Preferred Model		Without Transfers between NCR and Comparison Schools		Student FE		Remove all Movers and Leavers from NCR or Comparison Schools	
	(1) Math	(2) ELA	(3) Math	(4) ELA	(5) Math	(6) ELA	(7) Math	(8) ELA
Ever NCR*NCR Year	-0.009	-0.000	-0.008**	-0.000	-0.048***	0.005**	-0.054***	0.006**
	(0.005)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
Year 1	-0.017+	-0.006	-0.017+	-0.006	0.020**	0.004	0.024**	0.004
	(0.010)	(0.008)	(0.010)	(0.008)	(0.008)	(0.004)	(0.009)	(0.005)
Year 2	-0.006	0.009	-0.006	0.010	0.050**	0.023**	0.070***	0.018*
	(0.013)	(0.011)	(0.013)	(0.011)	(0.016)	(0.007)	(0.019)	(0.009)
Year 3	0.083+	-0.002	0.083+	-0.002	0.135***	0.025*	0.144***	0.020
	(0.042)	(0.015)	(0.042)	(0.015)	(0.037)	(0.012)	(0.041)	(0.014)
Ever NCR*Year 1	0.024+	0.012	0.024+	0.012	-0.025+	0.004	-0.019	-0.001
	(0.012)	(0.010)	(0.013)	(0.010)	(0.015)	(0.007)	(0.017)	(0.008)
Ever NCR*Year 2	0.113**	0.002	0.112**	0.001	0.089**	-0.003	0.072*	-0.006
	(0.040)	(0.012)	(0.040)	(0.012)	(0.031)	(0.009)	(0.034)	(0.011)
Ever NCR*Year 3	0.108	-0.013	0.109	-0.013	0.385***	0.001	0.379***	-0.002
	(0.095)	(0.017)	(0.096)	(0.017)	(0.042)	(0.013)	(0.047)	(0.015)
N (Student-Year)	492553	480134	489597	477366	387559	397233	295964	304225
R-Squared	0.788	0.888	0.787	0.888	0.766	0.957	0.771	0.959
Student Covariates	Yes	Yes	Yes	Yes	No	No	Yes	Yes
School FE	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Student FE	No	No	No	No	Yes	Yes	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note. Standard errors in parentheses. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. No lagged outcomes are included as covariates in the student fixed effect model. FE standards for fixed effect. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 17. Placebo Test for Effects in the Baseline Year before NCR Implementation

	(1) Math	(2) ELA
Ever NCR*NCR Year (Centered using Baseline Year)	-0.004	0.005
	(0.005)	(0.003)
Baseline Year	0.018	0.027**
	(0.011)	(0.010)
Ever NCR*Baseline Year	-0.020	-0.026
	(0.016)	(0.014)
N (Student-Year)	343928	335823
R-Squared	0.793	0.878
Student Covariates	Yes	Yes
School and Year Fixed Effect	Yes	Yes

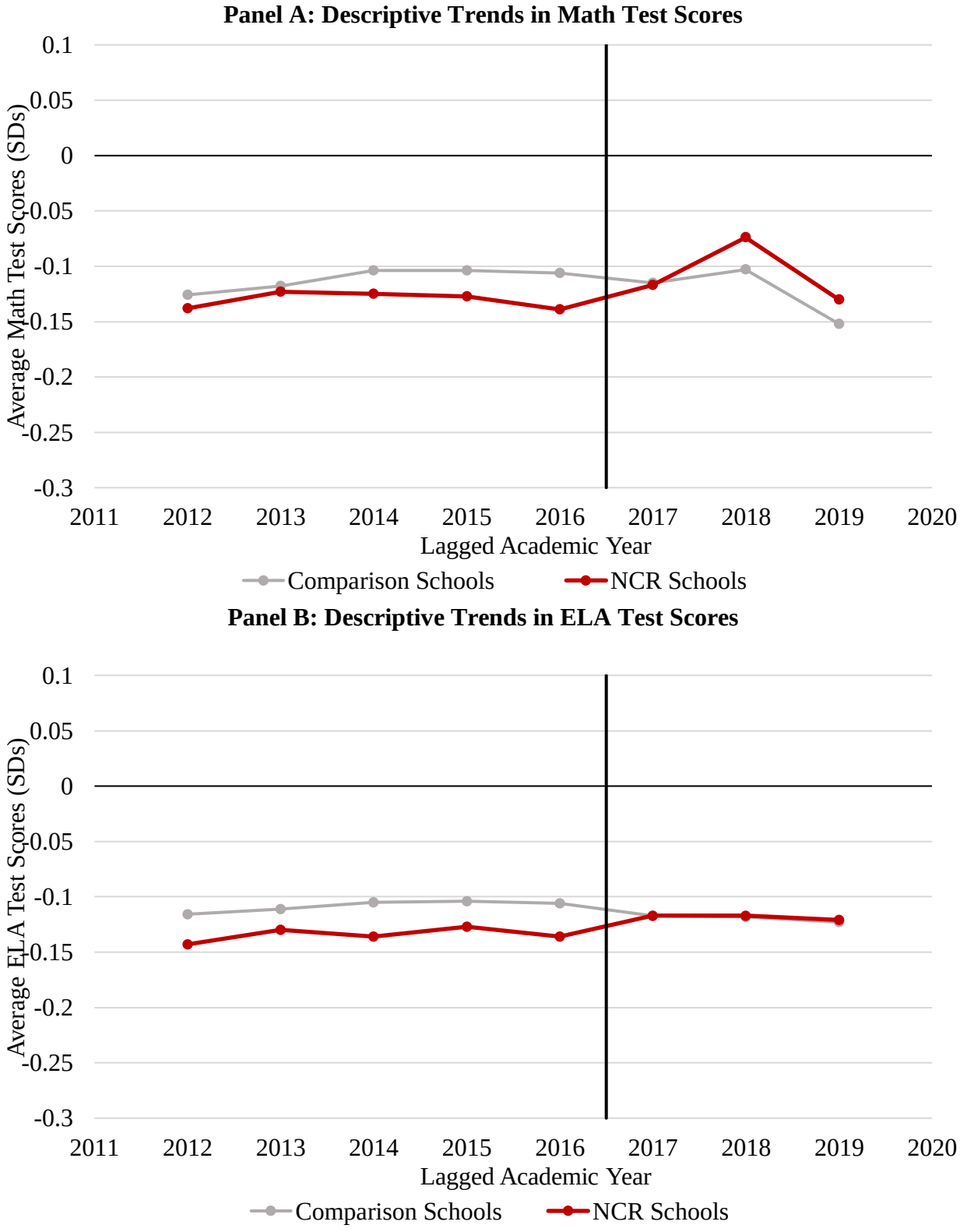
Note. Standard errors in parentheses. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 18. CITS Model with Alternative Samples

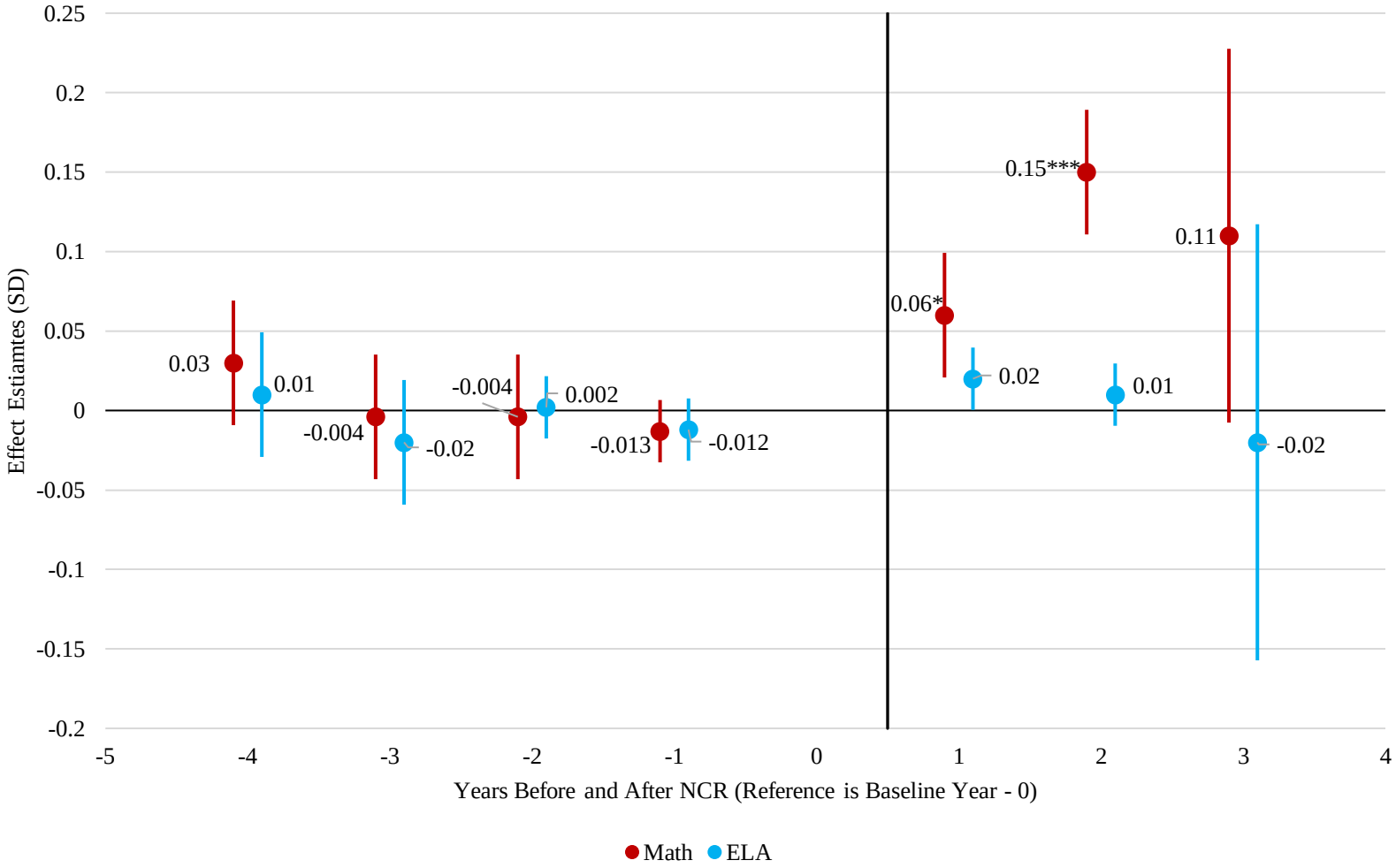
	Preferred Model		No ATR Schools		No Renewal Schools		Only Schools with Clear Start Date		Drop Schools that proposed School Choice	
	(1) Math	(2) ELA	(3) Math	(4) ELA	(5) Math	(6) ELA	(7) Math	(8) ELA	(9) Math	(10) ELA
Ever NCR*NCR Year	-0.007**	-0.000	-0.007*	-0.000	-0.008**	-0.000	-0.017	-0.005	-0.022+	-0.000
	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.028)	(0.005)	(0.012)	(0.002)
Year 1	-0.017+	-0.006	-0.016	-0.008	-0.018+	-0.005	-0.089*	-0.025+	-0.044*	-0.006
	(0.010)	(0.008)	(0.011)	(0.009)	(0.010)	(0.008)	(0.038)	(0.013)	(0.022)	(0.008)
Year 2	-0.006	0.009	-0.005	0.009	-0.008	0.010	-0.089	-0.016	-0.043	0.010
	(0.013)	(0.011)	(0.014)	(0.012)	(0.013)	(0.011)	(0.077)	(0.017)	(0.045)	(0.012)
Year 3	0.083+	-0.002	0.081+	-0.001	0.083+	-0.003	-0.013	-0.016	-0.040	0.002
	(0.042)	(0.015)	(0.043)	(0.015)	(0.042)	(0.015)	(0.159)	(0.022)	(0.107)	(0.015)
Ever NCR*Year 1	0.024+	0.012	0.023+	0.014	0.026*	0.012	0.085+	0.036*	0.038	0.012
	(0.012)	(0.010)	(0.014)	(0.011)	(0.013)	(0.010)	(0.048)	(0.015)	(0.037)	(0.011)
Ever NCR*Year 2	0.113**	0.002	0.118**	0.005	0.115**	0.002	0.185+	0.035+	0.101+	0.006
	(0.040)	(0.012)	(0.041)	(0.013)	(0.040)	(0.012)	(0.103)	(0.019)	(0.068)	(0.012)
Ever NCR*Year 3	0.108	-0.013	0.110	-0.013	0.108	-0.012	0.434**	0.011	0.382**	-0.012
	(0.095)	(0.017)	(0.096)	(0.017)	(0.095)	(0.017)	(0.158)	(0.026)	(0.118)	(0.017)
N (Student-Year)	492553	480134	451118	440094	486727	474297	213977	211799	454537	462282
R-Squared	0.788	0.888	0.791	0.891	0.786	0.887	0.524	0.892	0.567	0.889
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note. Standard errors in parentheses. ATR is the advanced teaching roles program. The NCR ever indicator and year trend main effects are omitted because they are perfectly collinear with the school and year fixed effects. All standard errors are clustered at the school level. Student covariates include indicators for gender, race, SWD, ED, ML status, and prior year lagged test score. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Figure 1. Descriptive Trends in Math and ELA Test Scores by Academic Year



Appendix Figure 2. Coefficients from More Parsimonious Event-Study with Each Year before and After NCR Began



Note. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ Model does not include Ever NCR * NCR Year or any fixed effects. Covariates include indicators for gender, race, SWD, ED, and ML status. Standard errors are clustered at the school level. Black vertical line demarcates the before and after NCR reforms began.