Accountability-Driven School Reform: Are There Unintended Effects on Younger Children in Untested Grades?

Abstract: Test-based accountability pressures have had mixed effects on the student outcomes that they are intended to improve. Accountability policies have also resulted in transfers of less effective teachers into untested early grades and more effective teachers in early grades into tested grades, which could yield unintended negative consequences. In this study, we use a sharp regression discontinuity design to examine the effects of an accountability-driven school reform on student outcomes and teacher mobility in 38 elementary schools assigned to reform in North Carolina. We find evidence of a small increase in chronic absenteeism and grade retention in grades K-2 in the first year of reforms. We also find suggestive evidence of negative effects on early literacy and reading comprehension, measured using formative reading assessments, in the first year that rebounded somewhat in the second year. Schools labeled low performing reassigned low effectiveness teachers from tested grades into untested early grades, though these assignment practices were no more prevalent in reform than control schools. Our results suggest that accountability-driven school reform can yield negative consequences for younger students that may undermine the success and sustainability of school turnaround efforts.

Keywords: school reform; test-based accountability; untested grades; early-grade outcomes; strategic staffing; teacher assignments

This is the final accepted unformatted version of Henry, G. T., McNeill, S. M., & Harbatkin, E. (2022). Accountability-driven school reform: Are there unintended effects on younger children in untested grades? *Early Childhood Research Quarterly*, 61, 190 -208. https://doi.org/10.1016/j.ecresq.2022.07.005.

Accountability-Driven School Reform:
Are There Unintended Effects on Younger Children in Untested Grades?

Test-based school accountability, which involves testing students, setting goals, and attaching consequences to failure to meet those goals, became ubiquitous under the No Child Left Behind Act [NCLB] (NCLB, 2001). The theory of change undergirding school accountability is that publicly labeling a school as low performing would motivate educators to improve their practice and ultimately lead to improved school performance. When NCLB did not bring the desired results through labeling and accountability alone, subsequent iterations of federal accountability policy (e.g., Race to the Top, School Improvement Grants, NCLB waivers) and some state policies aimed to supplement labeling with school supports and additional resources as well as requirements designed to improve student outcomes in low-performing schools. In implementing these policies, the federal government and some states invested heavily in turning around low-performing schools.

There is a substantial body of research investigating the effects of these reforms on intended outcomes (two recent meta-analyses describing this body of research are Redding & Nguyen, 2020; Schueler et al., 2021), which suggests modest positive effects on test scores in low-performing schools. But there are also concerns about unintended consequences of high-stakes accountability reforms that could thwart school improvement. Specifically, exams for test-based accountability only focus on a subset of grades and subjects, including federal requirements for testing reading and math in grades 3–8. The gaps or omissions in test-based accountability could create incentives for schools to focus their efforts and resources disproportionately on tested grades and subjects and away from untested early grades. For example, research has found that some schools—especially those that are low performing—

strategically place their most effective teachers in tested grades and subjects (i.e., grades 3 and above), and reassign less effective teachers to untested grades (i.e., grades K-2) in which they are less likely to influence a school's performance rating (Chingos & West, 2011; Cohen-Vogel, 2011; Fuller & Ladd, 2013; Goldring et al., 2015; Grissom et al., 2017; Kraft et al., 2020). Schools may also concentrate resources other than human capital in tested grades and subjects. Because spending and resources are critical to student and school success (Jackson et al., 2016), redirecting resources away from grades K-2 could negatively affect student learning and engagement. Reduced student learning and engagement as a result of redirecting resources away from untested early grades would undermine the goal of accountability-driven interventions such as school turnaround in low-performing schools. These unintended and possibly negative effects of accountability policies are especially concerning given that early childhood experiences have long-term effects on future outcomes, including subsequent achievement, college attendance, and earnings (Chetty et al., 2011; Dynarski et al., 2013; Schweinhart et al., 2005).

This study extends the literature on the impact of test-based accountability on early childhood investments by examining whether an initiative to turn around the lowest performing schools in North Carolina from 2015 to 2017 had unintended effects on student outcomes and assignment of teachers in untested grades. The state of North Carolina was an early adopter of school turnaround, which represents specific high-stakes accountability consequences in which schools that receive low accountability scores (measured primarily by state assessments) are provided with extra support and resources in order to improve school performance. All states under the Every Student Succeeds Act [ESSA] are required to implement these types of reforms in their lowest performing schools in some form. Reforms under ESSA are different from prior waves of federal turnaround initiatives, but align closely with the state of North Carolina's

turnaround intervention that we study here. To our knowledge, no prior study has examined the effects of school turnaround on early-grade student outcomes or whether turnaround schools are more likely to strategically redirect resources away from untested early grades. Thus, we draw from unique K-2 student achievement data that is not part of federal accountability testing to answer the following research questions:

1. What are the effects of efforts to improve the lowest performing schools, as identified by a test-based third- through eighth-grade accountability system, on student literacy and other student outcomes in untested grades?
2. Are the schools designated for turnaround by a state accountability system more likely to strategically assign teachers to and from untested grades based on teacher experience or effectiveness than other schools also labeled low performing?

We answer the first question using a sharp discontinuity design that compares outcomes for students in schools on either side of the eligibility cutoff for a school reform intervention, which assigned schools to receive turnaround services based on their school proficiency rates. We answer the second question using a descriptive analysis of staffing in turnaround and other low-performing schools in the state. The remainder of this paper proceeds as follows. In the next section, we overview the literature on the unintended effects of accountability policy on early grades and other subject areas. We then describe North Carolina's turnaround intervention, known as the North Carolina Transformation [NCT] initiative, and its associated theory of change. Next, we describe the study methods, including data, sample, measures and analysis for both research questions. Then we turn to results, followed by a discussion of findings, including relevance and implications for future accountability and school turnaround research.

Unintended Effects on Untested Grades

Early childhood education is critical to short- and longer-term outcomes—especially for students from disadvantaged backgrounds, English learners, and students that attend low-performing schools (Bassok, 2010; Currie, 2001; Lipsey et al., 2018; Weiland & Yoshikawa, 2013). Many studies find that students who participate in high quality early childhood programs—including preschool, pre-K, and kindergarten—benefit from improved outcomes from childhood into adulthood, including higher student achievement, socioemotional development, high school and college completion rates, and adult earnings, as well as lower rates of criminal activity (Atteberry et al., 2019; Chetty et al., 2011; Deming, 2009; Dynarski et al., 2013; Johnson & Jackson, 2019; Schweinhart et al., 2005). However, other studies find fade-out of short-term gains from preschool and pre-K (Li et al., 2020; Phillips et al., 2017) or even reversal of initial positive effects (Lipsey et al., 2018). While the effects of pre-K have been studied extensively, a recent consensus panel suggests that the quality of early elementary grade experiences is critical to whether children can sustain or even amplify early learning gains (Phillips et al., 2017). Collectively, current research suggests that lowering the quality of early elementary experiences through strategies such as systematically assigning less effective teachers to early grades could negatively affect student learning and longer term outcomes.

However, shortly after states began to invest in early childhood programs in the 1990s, No Child Left Behind ushered in a high-stakes accountability era that incentivized states to prioritize student performance on standardized tests. Specifically of relevance to achievement in early elementary grades, test-based accountability programs typically assess school performance based exclusively on student achievement on standardized tests in third through eighth grades. In order to boost school performance and avoid the consequences of being labeled a “failing”

school, educational leaders may respond to these accountability pressures by disproportionately concentrating resources in tested grades and subjects and away from untested early grades. Theories based on numerous social science disciplines have contributed to our understanding of possible mechanisms through which test-based accountability may produce unintended consequences. Insights from economics suggest that test-based accountability can lead to "perverse incentives" for schools to raise proficiency rates by pushing out or suspending lower performing students, triaging resources to students at the cusp of state proficiency levels to the detriment of students throughout the test score distribution, classifying low-performing students as having disabilities, or engaging in other gaming behaviors including outright cheating (Ballou & Springer, 2016; Booher-Jennings, 2005; Cohen-Vogel, 2011; Ho, 2008; Jacob & Levitt, 2003; Ladd & Lauen, 2010). Both economists and sociologists have focused on strategic staffing—economists from the perspective of reallocation of resources and sociologists from a lens of organizations and staffing. Sociology provides insights through which to understand the extent to which staffing decisions may affect educational opportunities and ultimately reduce or exacerbate achievement gaps (Gamoran, 1987; Kalogrides et al., 2013). Political science and policy research elucidates the ways that micropolitical dynamics can contribute to teacher classroom assignments (Clotfelter et al., 2006; Grissom et al., 2015).

Teachers are among the most influential resources through which schools can influence student outcomes (Aaronson et al., 2007; Adnot et al., 2017; Ladd & Sorensen, 2017; Rockoff, 2004), and schools may reallocate teaching resources through a practice that is known as strategic teacher assignment. Specifically, some principals report strategically assigning higher quality teachers based on teacher effectiveness data to tested grades and subjects because performance in these grades and subjects counts toward school performance scores and

designations. In turn, these principals report "hiding" ineffective teachers in untested grades and subjects in which they are less likely to influence a school's performance rating (Cohen-Vogel, 2011; Goldring et al., 2015). Large-scale, quantitative studies of teacher assignments in Florida and North Carolina provide empirical evidence that schools assign more effective and highly qualified teachers to high-stakes tested grades or subjects, and less effective and less highly qualified teachers to low-stakes early grades (Chingos & West, 2011; Fuller & Ladd, 2013; Grissom et al., 2017; Kraft et al., 2020). Fuller and Ladd (2013) found that North Carolina elementary schools under NCLB were more likely to move plausibly higher quality teachers up to tested grades (3-5) and lower quality teachers down to untested grades (K-2), where quality is measured by Praxis exam scores and other credentials, including experience. Kraft and colleagues (2020) found similar strategic staffing patterns based on principal performance ratings of teachers in one large North Carolina district between 2002 and 2010. These practices are especially prevalent in schools with low accountability grades or labels, likely due to increased pressure to improve school performance (Chingos & West, 2011; Cohen-Vogel, 2011; Fuller & Ladd, 2013; Goldring et al., 2015; Grissom et al., 2017).

Considering the evidence on the importance of early childhood experiences for future life outcomes (Atteberry et al., 2019; Chetty et al., 2011; Dynarski et al., 2013; Johnson & Jackson, 2019; Schweinhart et al., 2005), the practice of concentrating resources in tested grades and subjects in response to accountability pressures may have unintended negative consequences in early grades that spill into later achievement. Indeed, Grissom and colleagues (2017) linked strategic staffing practices to unintended effects on early-grade student achievement. The authors found that the reassignment of less effective teachers to untested grades led to lower early-grade

student achievement gains—measured by the low-stakes Stanford Achievement Test—and that these losses persisted into tested grades.

Student achievement in early grades may also suffer from the stigma of the low-performing label (Finnigan & Gross, 2007). Qualitative evidence suggests that the low-performing label attaches a stigma to schools for teachers and students and can lead to demoralization that may in turn increase absenteeism and reduce achievement outcomes (Murillo & Flores, 2002; Rice & Malen, 2003). Additionally, the turnaround reform itself may erode trust and lead to stigma and demoralization among staff and the broader school community (Maxcy, 2009). Teachers in early grades, in particular, may respond differently to the low-performing label and the reform because they are subject to the stigma without the benefits of instructional supports and additional resources that are typically targeted at teachers of tested grades. Indeed, several studies suggest that demoralization arising from school reforms can undermine improvement efforts (Hamilton et al., 2014; Hess, 2003; Mathis, 2009). To that end, it is possible that the low-performing designation may undercut morale, teaching effort, and instructional quality in early grades by introducing a stigma or atmosphere of distrust without counteracting that stigma with supports.

While a plethora of research has examined the effects of school accountability on teacher assignments and student outcomes, no research to date has done so as a response to school turnaround—a prevalent policy since Race to the Top that all states are required under ESSA to implement in their lowest performing schools. A paradox of school turnaround is that its theory of change calls for improving the systems underlying school performance but the accountability mechanisms in place focus only on a subset of grade levels and subject areas. In order to truly improve the lowest performing schools, it is therefore critical to understand the extent to which

turnaround impacts school practices and student achievement in areas not subject to the accountability spotlight. While research has shown substantial heterogeneity in the effects of school turnaround on student outcomes in grades 3 and above (Carlson & Lavertu, 2018; Dougherty & Weiner, 2017; Henry & Harbatkin, 2020; Pham et al., 2020; Strunk, Marsh, Hashim, et al., 2016; Zimmer et al., 2017), this is the first to our knowledge to examine the effects of a turnaround intervention on student outcomes and staffing in untested early grades.

Our research aims are as follows. We examine the effects of accountability-driven school reform within the context of the North Carolina Transformation (NCT) initiative, which designated the lowest performing schools—based on standardized test score proficiency in 2015—as turnaround schools. Specifically, we assess the effects of NCT on early grade literacy and other student outcomes by comparing student outcomes in NCT schools to those in a set of similarly low-performing control schools. Literacy outcomes include early literacy skills and text reading comprehension—measured using the mCLASS Dynamic Indicators of Basic Early Literacy Skills [DIBELS] and Text Reading and Comprehension [TRC] assessments, respectively—and other student outcomes include chronic absenteeism and grade retention. Lastly, we assess whether low-performing turnaround schools are more likely to strategically assign teachers to tested and untested grades based on teacher experience or effectiveness than other schools labeled low performing, a pattern previously documented in North Carolina during the NCLB era (Fuller & Ladd, 2013; Kraft et al., 2020).

North Carolina Transformation Initiative

The North Carolina Transformation [NCT] school turnaround initiative was implemented in 75 low-performing schools across the state during the 2015-16 and 2016-17 school years. Thirty-five of these 75 schools enrolled K-2 students, the grades that are the focus of the present

study. We note that while including first- and second-grade students may extend the period of early childhood beyond the ages of children many studies include in early childhood, children in early elementary grades may reasonably be considered in their early childhood years. There is sparse research on the effects of accountability-mandated reforms on the young children in grades K-2, and these grades have important implications for the "fade out" of effects of earlier childhood interventions. NCT served the state's low-performing schools during the period between Race to the Top and ESSA, and the NCT model aligns closely with ESSA's flexible approach to school turnaround. The intervention was overseen by the North Carolina Department of Public Instruction under the direction of their District and School Transformation unit. Figure 1 graphically displays the theory of change for the NCT intervention, developed by the District and School Transformation unit at the outset of the intervention.

Figure 1 ABOUT HERE

After a school received the NCT designation, the intervention design called for services to begin with a comprehensive needs assessment in which coaches assigned by the District and School Transformation team reviewed school achievement data; interviewed principals; held focus groups with school staff, students, and parents; and conducted classroom observations in treatment schools to identify the strengths and weaknesses of the school and assess where supports should be targeted. Similar needs assessments are required for all low-performing schools under ESSA, which is grounded in evidence suggesting that identifying school needs is a critical early step toward school improvement (Herman et al., 2008; Wallace Foundation, 2010). Due to resource constraints, comprehensive needs assessments occurred throughout the course of the intervention rather than at the outset, and turnaround supports began in many schools prior to the comprehensive needs assessments being carried out (Henry & Harbatkin, 2020).

Comprehensive needs assessment findings were then "unpacked" or discussed with treatment school staff. These 1.5-day unpacking sessions involved reviewing the comprehensive needs assessment findings, conducting a "root cause analysis" that identified the causes underlying issues at the school, and conducting a "brown paper planning" activity that visually displayed the school improvement process. Unpackings generally occurred during the summer following the school year of the comprehensive needs assessment, although there was variation in when and whether schools received unpackings.

Following the comprehensive needs assessment and unpacking, the theory of change called for schools to create their school improvement plans to outline their priorities and goals. Schools then submitted school improvement plans through an online platform called NCStar, and state coaches provided feedback through the same platform. Evidence suggests that a comprehensive planning process that draws on needs assessment data to develop plans such as the process outlined in the theory of change is an integral component of successful turnaround (Herman et al., 2008; Meyers & Hitt, 2018; Strunk, Marsh, Bush-Mecenas, et al., 2016). The comprehensive needs assessment, unpacking, and school improvement plan were ostensibly focused on the whole school rather than exclusively on tested grades.

The core of the intervention was the coaching that followed. Based on the comprehensive needs assessment, unpacking, and school improvement plan, coaches were assigned to NCT schools with the goal of building school capacity. School transformation coaches worked with principals and instructional coaches worked with teachers. Nationally, coaching is a focus of school turnaround policy, and has in some contexts led to student achievement gains (Kraft et al., 2018). Coaching can also increase buy-in to a turnaround intervention or direct efforts toward particular priorities (Coburn & Woulfin, 2012; Woulfin, 2015). In many states as in North

Carolina, assigning coaches to low-performing schools is a component of turnaround interventions (Meyers & VanGronigen, 2018; VanGronigen & Meyers, 2019). Under NCT, there were no formal or state-mandated coaching requirements; instead, coaches provided tailored supports to principals and teachers based on their needs. Over the three semesters of coaching from spring 2016 through spring 2017, schools assigned to treatment received an average of 37 instructional coach visits and 19 school transformation coach visits. However, there was large variation in the number and content of coaching visits by school, with instructional coach visits ranging from 0 to 79 and school transformation coach visits ranging from 0 to 49. Qualitative evidence and conversations with personnel implementing the coaching suggest this variation stemmed from factors related to tailoring the coaching to meet school needs rather than a prescribed, predetermined number of visits or state capacity to deliver services due to budget constraints rather than school or district leadership decisions to take up the services (Herman et al., 2019).

Because coaching visits were aligned with the school improvement plan, they were likely to be concentrated in the tested grades and subjects because the performance measure was performance on third- through eighth-grade high-stakes tests in reading, mathematics and science. Thus, the intervention was intended to improve the whole school but targeted specific, tested grade levels and subjects. Teachers in untested early grades were subject to the disadvantages of the intervention—the low performing and turnaround labels along with any demoralization associated with the needs assessment findings—but not the potential benefits of the coaching. In particular, educators in NCT schools reported that the low-performing label created a stigma that created new challenges around teacher recruitment and retention and parent and community engagement (Marks & Holly, 2019). Based on the theory of change, the planning

along with school transformation and instructional coaching was expected to lead to changes in principal and teacher practices, outcomes, and retention. In turn, student outcomes were expected to improve.

In particular, the intervention was focused on improving proximate student outcomes such as attendance, on-time grade progression, and behavior, and more distal outcomes such as student achievement in tested grades and subjects. Another study examined these intended effects NCT and found that the intervention did not produce the desired intended effects on teacher or student outcomes. Specifically, Henry & Harbatkin (2020) examined the effect of NCT on student test score growth on end-of-grade and end-of-course exams in grades 4 and above found no effect in the first year of the intervention followed by a 0.13 standard deviation decline in test score growth and a 22 percentage point increase in teacher turnover in the second year. The negative effects appeared to be associated with the timing and nature of the comprehensive needs assessments that were delivered (Henry & Harbatkin, 2020). To that end, negative effects of the intervention may have extended to untested grades, and may be even larger in these grades if turnaround schools strategically reassigned less experienced and/or effective teachers from tested to untested grades. While the prior study focused only on intended effects in tested grades (i.e., whether NCT improved outcomes targeted in the theory of change), this study adds a further contribution to the literature by investigating the effects in untested grades. This contribution includes examinations of literacy and other outcomes in early grades and of unintended effects in the form of strategic staffing.

Methods

Data and Sample

This study relies on two sources of data. We draw from statewide administrative data from a longitudinal database maintained by the University of North Carolina-Chapel Hill's Educational Policy Initiative at Carolina containing data on all students, teachers, and schools in North Carolina. We use data from 2014-15 through 2017-18. We merge the student administrative data with mCLASS K-2 student literacy data from the 2015-16 and 2016-17 school years. We draw from the student-level data to answer our first research question, using student-by-testing period (beginning of year and end of year, respectively) data to examine effects on student literacy and student-by-year data to examine effects on chronic absenteeism and grade retention. To answer the second research question, we merge teacher experience and evaluation data with student course-level data to understand the extent to which teachers with different effectiveness and experience levels are assigned to different grade levels.

The full sample includes the 175 North Carolina schools that enrolled K-2 students in both the 2016 and 2017 school years and were eligible for treatment under NCT. Schools were excluded from NCT eligibility if they had a school performance grade of C or above for the 2014-2015 school year, exceeded expected growth, were part of one of the 10 largest school districts in the state or in Halifax County (which participated in a separate, district-level turnaround during the same time as the NCT intervention), or were designated as a special or charter school.

The state assigned schools to participate in the NCT intervention based on their 2014-15 school performance composite, a measure that represents grade-level proficiency on state assessments in grades 3 and above. In the study sample, these assessments exams include third through eighth grade math and reading, and fifth and eighth grade science. The cutoff score for NCT participation was 31.1 for schools enrolling K-2 students, with 38 schools scoring below

31.1 being targeted for services and 35 actually receiving services. Before beginning turnaround services, the state sought permission from districts. In a few instances, district officials requested substitution of a school above the threshold receive services for, or in addition to, a school below the threshold. As a result, 32 of the 38 schools below the threshold received NCT services, six below the threshold declined services, and three above the threshold received services. Eligibility for NCT was a strong predictor of participation in NCT, as we show visually in Appendix A (Figure A.2). Schools below the cutoff value of zero had a high probability of participation in the NCT intervention, whereas schools above the cutoff had a low probability of participation. In total, 38 schools that enrolled K-2 students were assigned to treatment as a result of having proficiency rates below the threshold, and 35 schools actually participated in the NCT intervention. The sharp RD design assigns to treatment those 38 schools below the cutoff. The control group comprises other low-performing schools with proficiency rates just above the 31.1 threshold.

Baseline school-level sample characteristics are displayed in Table 1. Following What Works Clearinghouse standards for RD designs (What Works Clearinghouse, 2020), we calculate each of these values using a sharp RD with the same functional form as our main models (described in detail in the statistical analysis strategy section below) but with the variable listed in each row as the outcome variable. Specifically, What Works Clearinghouse requires that conditional on the forcing variable, there is no impact of the intervention on baseline covariates at the cutoff. There are no significant differences in student demographics, teacher demographics, or school performance between treatment and control schools, controlling for the forcing variable.

Table 1 ABOUT HERE

The student sample includes 49,017 unique students who were in K-2 during the study period from 2015-16 through 2016-17. The teacher sample includes 5,126 unique teachers of grades K-2 and tested subjects in grades 3-8 who taught in a treatment or control school beginning with the year prior to the study period (i.e., 2014-15 through 2016-17) in order to examine teacher pathways into and out of untested lower grades.

Research Design

We use a regression discontinuity [RD] design to estimate the effect of the intervention on student literacy and other outcomes. The RD provides a local average treatment effect of NCT for students in schools near the eligibility cutoff described above. To obtain a causal estimate of NCT on student outcomes, we leverage the fact that schools were assigned to NCT based on their 2014-15 proficiency rate—comparing students in schools just below the proficiency rate cutoff with their peers in schools just above the cutoff.

Measures

In this section, we begin by describing our student literacy outcomes, including early literacy and reading comprehension, from which we draw to answer our first research question. We turn next to other student outcomes, including chronic absenteeism and grade retention, which also support our first research question. We then move to our teacher assignment outcomes, which we use in the descriptive analysis to answer the second research question.

Student Literacy Outcomes

We estimate the effects of NCT on two student literacy outcomes: early literacy and reading comprehension, which are measured using the mCLASS Dynamic Indicators of Basic Early Literacy Skills [DIBELS] and Text Reading and Comprehension [TRC] assessments, respectively. The DIBELS assessment is composed of multiple, one-minute subtests of student

phonemic awareness, alphabetic knowledge, and reading and retell fluency (the specific subtests on which students are assessed varies by grade and time of school year; for further details, see Good & Kaminski, 2002). The TRC assessment assesses reading accuracy, fluency, and comprehension through having students read leveled benchmark books and completing follow-up comprehension tasks. Both assessments are administered three times per school year, at the beginning, middle, and end of the school year. While the mClass is intended as a formative assessment, validation research has shown that both the DIBELS and TRC assessments have high validity and reliability (Amplify Education, 2014; Good & Kaminski, 2002; Smith et al., 2020). At the time of our study, only early literacy and reading comprehension was assessed statewide using the mCLASS assessment in North Carolina in grades K-2. This prevented us from incorporating measures of K-2 student performance in other subject areas, such as math and science, in this study.

We operationalize early literacy as the end-of-year composite score from the mCLASS DIBELS early literacy assessment and reading comprehension as the end-of-year composite score from the mCLASS TRC reading comprehension assessment. We standardize the DIBELS and TRC composite scores by grade, year, and period (i.e., beginning- or end-of-year exam) to have a mean of zero and a standard deviation of one. Thus, for example, a standardized DIBELS end-of-year composite score of 0.025 denotes that a student performed 0.025 standard deviations above average on the end-of-year assessment relative to other students in their grade and year that were included in the study sample.

Other Student Outcomes

We also estimate the effects of NCT on two other student outcomes: chronic absenteeism and grade retention. Chronic absenteeism is operationalized as a binary indicator that takes a

value of 1 when a student is absent for 10 percent or more of enrolled school days, in line with the state of North Carolina's definition of chronic absenteeism and with other studies of chronic absenteeism (Gottfried & Hutt, 2019). The majority of states that include chronic absenteeism in their accountability formulas under ESSA also use a similar operationalization (Jordan & Miller, 2017). We operationalize grade retention as a binary indicator that takes a value of 1 for students who are retained in the same grade for a second year. Grade retention is measured at the end of the school year, so a student who repeats kindergarten in 2016-17 would be coded as being retained in 2015-16.

Table 2 provides student-level descriptive statistics for both sets of outcomes, first for the full sample and then within the optimal bandwidth (which we describe in the statistical analysis strategy section below). We highlight that due to the RD design, assignment to treatment is effectively random and the treated and control schools within the optimal bandwidth (columns 3 and 4) were similar at the time of treatment assignment. These descriptive differences in Year 1 and 2 of the intervention therefore point to post-treatment differences in means but these should not be interpreted as causal estimates because the means are not adjusted for the forcing variable; the RD results presented in Table 3 below provide estimates of the treatment effects.

Table 2 ABOUT HERE

Teacher Assignments, Experience, and Effectiveness

Our second research question examines whether NCT schools are more likely to strategically reassign teachers to untested early grades based on teacher experience or effectiveness. To examine teacher assignments, we draw from teacher experience and evaluation data merged with student course-level data. We code a teacher as teaching in a tested grade and subject if she teaches a grade-subject combination with an end-of-grade exam. In this sample,

teachers of tested courses include those teaching math or reading to students who take math or reading end-of-grade exams in third through eighth grade, or who teach science to students who take science end-of-grade exams in fifth or eighth grade. We code a teacher as teaching in an untested early grade if she teaches only students in untested early grade academic grades and subjects. In this sample, a teacher would be coded as teaching an untested early grade if she teaches K-2 math, science, reading, or social studies and she is not coded as also teaching in a tested grade or subject.

To examine the role of teacher experience on teacher assignment, we classify teachers as novice if they have fewer than four years of experience, in line with the state's definition of novice teacher and with other studies of teacher experience (e.g., Araujo et al., 2016; Glennie et al., 2016; Graham et al., 2020). Similar to other studies of strategic staffing (Chingos & West, 2011; Fuller & Ladd, 2013; Grissom et al., 2017), we measure teacher effectiveness using subject-level value-added scores (specifically, the Education Value-Added Assessment System, or EVAAS). The state calculates EVAAS scores using end-of-grade exams for teachers in tested courses and using mCLASS TRC reading comprehension assessments for K-2 teachers, whose students do not take end-of-grade exams. EVAAS scores are a continuous measure that can theoretically range from negative to positive infinity (Wright et al., 2010). Across all teachers in the state, the mean EVAAS score is zero and the standard deviation is two. EVAAS scores are available for about 90 percent of teachers in the sample.

Teachers receive one of three ratings from the state based on their EVAAS score for a given subject— *exceed expected growth* if they have a EVAAS score greater than +2, *do not meet expected growth* if they have a EVAAS score of less than -2, and *meet expected growth* if their EVAAS score falls within two points of the mean (i.e., between -2 and +2). We therefore

follow these growth ratings set by the state and received by school leaders, coding a teacher as "low effectiveness" if they are rated as not meeting expected growth, "high effectiveness" if they are rated as exceeding expected growth, and "mid effectiveness" if they are rated as meeting expected growth. Across our full sample of treatment and control schools, about 20 percent of teachers with EVAAS scores are low effectiveness, 66 percent are mid effectiveness, and 14 percent are high effectiveness. These categorical ratings are particularly salient because the state provides them to principals making staffing decisions for the new school year. Thus, EVAAS scores are one of the few measures of teacher effectiveness available to principals in the state of North Carolina when making teacher assignments.

While EVAAS scores capture only one dimension of teacher effectiveness (i.e., teachers' contributions to student academic learning), value-added scores have been shown to be valid measures of effectiveness that identify teachers who produce higher achievement among their students (Kane et al., 2013). Our analyses in particular draw from averages (do schools assign teachers classified as highly effective to different classes than teachers classified as low effectiveness?)—while EVAAS may not perfectly capture each individual teacher's effectiveness level, the teachers who receive a high effectiveness rating are, on average, more highly effective than the teachers who receive a low effectiveness rating at least on this particular dimension of teacher effectiveness.

Controls

We include a robust set of school, teacher, and student covariates. School-level covariates include minority percentage, economically disadvantaged percentage, per-pupil expenditures (PPE) and PPE squared, and enrollment and enrollment squared. Teacher-level covariates include gender and race/ethnicity with white as the reference category. Student-level covariates

include grade level with kindergarten as the reference category, female, race/ethnicity with white as the reference category, disabled, limited English proficient (LEP), over-age for grade, and nonstructural transfer in. We define disabled as currently designated with any exceptionality code other than academically gifted. We define over-age as having a birthdate that would place the student in a grade level above the grade level assigned. We define nonstructural transfers in as transfers that occur into the observed school after the beginning of kindergarten. We also include four additional student-level variables in our models that measure variation in the administration of the mClass assessments: beginning-of-year early literacy or reading comprehension score, a dichotomous variable indicating whether the student was assessed by their own classroom teacher at beginning of the school year, a dichotomous variable indicating whether the student was assessed by their own classroom teacher at end of school year, and days between beginning- and end-of-year assessments. During the study period, state policy allowed for either teachers or external assessors to administer the mClass assessments. The DIBELS and TRC beginning-of-year assessments in grades K-2 were supposed to be given by the classroom teacher so that the teacher could use the results to guide personalized instruction. A certified staff member was supposed to assess students in TRC at the end of the year, whereas the classroom teacher could continue to assess students using DIBELS. The beginning-of-year mCLASS exams are administered within the first 25 days of the school year and end-of-year exams within the last 30 days of the school year.

Statistical Analysis Strategy

Regression Discontinuity Design

We estimate the effect of being just below the threshold for assignment to NCT on K-2 student outcomes using a regression discontinuity [RD] design, which exploits the jump in

probability of assignment to treatment at the treatment eligibility cutoff (Imbens & Lemieux, 2008). This approach allows us to estimate the effect of assignment to treatment for schools around the cutoff, or the local average treatment effect. As long as the score on the assignment variable and threshold for eligibility are exogenously determined, assignment to treatment or control is considered effectively random around the cutoff. In this case, the state set the eligibility threshold based on available resources; they wanted to serve 75 total schools and they wanted half of those to be elementary schools because elementary schools comprise half the schools in the state. We therefore have no evidence that the state manipulated the cutoff—a critical assumption for the validity of the RD design that we explore later.

To model the effect of NCT around the cutoff, we estimate regression models with the student outcomes (i.e., early literacy, reading comprehension, chronic absenteeism, or grade retention) as outcomes and including the assignment to treatment indicator and a flexible function of the 2014-15 school proficiency rate (the forcing variable) that can vary on either side of the cutoff. To calculate the optimal bandwidth around the cutoff—that is, the maximum distance from the eligibility threshold on which we will compare outcomes—we use the mean square error optimal bandwidth selection procedure described by Calonico, Cattaneo, & Titiunik (2014), which accounts for the clustered assignment of schools to treatment. The model also includes vectors of school- and student-level covariates described in the Measures section, including the student's score on the beginning-of-year exams for early literacy and reading comprehension depending on the outcome being analyzed. Moving forward, we abbreviate the bandwidth selected using this procedure as the "CCT" bandwidth.

We estimate heteroskedasticity-robust standard errors rather than clustering at the school level because the relatively few number of clusters may lead cluster-robust standard errors to

provide a biased estimate of the true variance and over reject the null (Cameron & Miller, 2015). We also estimated a set of models with standard errors clustered at the school level to account for correlated errors within schools and obtained similar results. These results, like multilevel models with students nested in schools, account for correlated errors within units. We choose to present the results using heteroskedasticity-robust standard errors because they provide a more conservative hypothesis test.

Because the literacy outcomes are standardized, the estimated effects from these models can be interpreted as effect sizes in standard deviation units. Because the chronic absenteeism and grade retention outcomes are binary indicators, models predicting these outcomes are linear probability models, which means the estimated treatment effect in these models represents the difference in probability of chronic absenteeism or grade retention for students in NCT schools relative to students in control schools. We estimate all models separately for each year of treatment.

The resulting effect estimates are intent-to-treat (ITT) estimates because they capture the effect of being assigned to treatment, regardless of treatment take-up. We consider the ITT approach to provide the policy-relevant estimates, which are the estimated effects for state policymakers considering a similar policy because policy decisions should not assume ubiquitous take-up of treatment. The ITT estimates are not subject to bias arising from differences between schools that complied or did not comply with their original treatment assignment. Intuitively, the RD model assumes that schools right around the treatment cutoff are equal in expectation and that any differences, conditional on the forcing variable, can therefore be attributed to the treatment. In other words, the treated schools right near the eligibility cutoff (i.e., the 14 schools within 2.9 points below the eligibility cutoff of 31.1) would fare similarly to

the untreated schools right near the cutoff (i.e., the 12 schools within 2.9 points above the cutoff) in the absence of treatment. Thus, we can assume that any differences between those two sets of schools are a result of NCT.

The validity of the RD estimator relies on several assumptions, including that there was no manipulation of the forcing variable (i.e., the value of the 2014-15 school performance composite was not manipulated to influence treatment assignment) and that the functional form of the relationship between the outcome and forcing variable is correctly specified. To examine the validity of these assumptions, we follow the What Works Clearinghouse guidelines (2020) for RD designs. We find the validity of the assumptions for the RD design is supported. Due to the limited number of schools within the optimal bandwidth, we also estimate the effect of NCT using a local randomization RD design (Cattaneo et al., 2015, 2016) as an additional validity check. This process involves identifying windows within which the sample is well balanced on baseline covariates on either side of the cutoff, calculating the mean difference within the balanced windows, and calculating p-values for those estimates under finite-sample assumptions (Cattaneo et al., 2016). For further discussion of the RD assumptions and results of all validity checks, see Appendix A.

To test the sensitivity of our results and meet additional What Works Clearinghouse standards, we run models within a series of alternative bandwidths, including 150% and 200% of the CCT bandwidth. We do not estimate on 50% of the CCT bandwidth because the bandwidth size includes only five schools below the cutoff and seven schools above the cutoff. We also estimate separate models by grade level, which can be found in Appendix B.

Strategic Staffing Logistic Regression

To answer our second research question, we compare assignment of teachers to untested grades separately based on teacher effectiveness (in one model) and experience (in another model). Specifically, we are interested in whether less effective and less experienced teachers are more likely to be assigned to untested courses, which would potentially lead to reduced learning for younger students. Our analytic sample for this analysis comprises teachers of both tested and untested grades in treatment and control schools during the study period (t = 2015, 2016, and 2017). We do not include teachers of untested non-academic subjects (e.g., physical education, music) in in our analysis. To classify teachers as effective or ineffective, we need teachers to teach either an end-of-grade exam course (i.e., 3-8 reading or math, 5 or 8 science) or K-2 reading, for which teachers receive EVAAS scores based on their students' mClass reading comprehension scores. Therefore, we begin with two samples in each year. The first sample comprises all tested teachers. These teachers receive EVAAS scores based on end-of-grade exams. The second sample is comprised of teachers who teach early grade reading (i.e., reading in K, 1, or 2). These teachers receive EVAAS scores based on mClass reading comprehension exams.

Using logistic regression, we separately estimate the logged odds that tested and untested teachers return to the same school and teach in an untested early grade in year t+1. Teachers who return to the same school and teach in an untested early grade in year t+1 are coded as 1 for the dichotomous outcome, while teachers who either (a) return to the same school and teach in a tested course in year t+1, (b) return to the same school and teach only untested courses (e.g., physical education or music) or (c) leave the school, are coded as 0. We estimate these logged odds ratios in separate equations for teachers of tested grades and subjects in year t and for

teachers of untested early grades (K-2 reading) in year *t* to account for differences in the probability of effectiveness classification in formative and accountability-based exams. Teachers are more likely to be classified as effective using TRC scores than end-of-grade exam scores, so teachers who are in untested grades in year *t* are disproportionately classified as highly effective relative to teachers in tested grades and subjects.

We run two sets of these models, with the first estimating the logged odds of teacher assignment using teacher effectiveness based on EVAAS scores, and the second estimating the logged odds of teacher assignment using teacher experience. Specifically, we classify teachers as high, mid, or low effectiveness based on their prior EVAAS score, and as experienced (4+ years of experience) or novice (fewer than 4 years of experience), respectively. We then predict the dichotomous outcome of returning to the same school in an untested early grade in year $t+1$ (relative to either leaving the school or returning and teaching outside K-2 academic subjects in year $t+1$) as a function of treatment assignment, teacher effectiveness category (*Low* and *High*, with mid-effectiveness as the omitted category), interactions between treatment and effectiveness category, vectors of school and teacher-level covariates, and an idiosyncratic error term clustered at the school level. We focus on this outcome because the teachers who end up in untested grades are of critical importance to younger students' learning; therefore, strategic staffing practices that reassign highly effective teachers away from these early grades or attempt to hide ineffective teachers in these early grades have potential negative implications for early grade student achievement.

Evidence of strategic staffing across the entire study sample is gaged by the coefficients on the low effectiveness and high effectiveness variables, while evidence of differential strategic staffing practices in treatment schools is determined by the coefficients on the interactions

between these variables and the treatment indicator. A positive estimate on the low effectiveness coefficient would provide evidence of strategic staffing with respect to low effectiveness teachers, while a negative estimate on the high effectiveness coefficient would provide evidence of strategic staffing with respect to high effectiveness teachers. In particular, a positive estimate on the low effectiveness main effect would suggest that low-effectiveness teachers were more likely to return to the same school and teach in an untested early grade across the full sample, and a positive estimate on the interaction between the low effectiveness main effect and the treatment indicator would suggest that strategic assignment of low effectiveness teachers to untested grades was more prevalent in NCT schools than control schools. A negative estimate on the high effectiveness coefficient would suggest that highly effective teachers were less likely to return to the same school and teach in an untested early grade across the full sample of schools, while a negative estimate on the interaction between the high effectiveness main effect and the treatment indicator would suggest that assignment of highly effective teachers to untested grades was less prevalent in NCT than control schools. If schools were not engaging in strategic staffing to the potential detriment of untested early grades, we would not expect to see significant estimates on each of these coefficients. We estimate parallel models for teacher experience in which we replace the low, mid, and high-effectiveness indicators and interactions with indicators that take the value of 1 for experienced teachers. We report these coefficients as odds ratios in which an estimate of greater than 1 indicates that the group of teachers is more likely to teach in an untested grade relative to the omitted reference group (mid-effectiveness teachers in the teacher effectiveness models and novice teachers in the teacher experience models), and a value below 1 indicates the group is less likely to teach in an untested grade than the omitted reference group.

Results

The results section proceeds as follows. We first discuss the effects of the NCT intervention on student literacy, followed by the intervention effects on other student outcomes. We then describe our findings on the strategic reassignment of teachers in untested early grades. The effects discussed below, which were estimated using a RD design, can be interpreted as effect sizes in standard deviation units.

Research Question #1: Effects of NCT on Literacy and Other Student Outcomes

Student Literacy Outcomes

We find evidence that NCT produced negative effects on early literacy and reading comprehension in the first year of the intervention followed by positive effects in the second year. We show these results graphically without controls in Figure 2. The first row displays results in early literacy and the second in reading comprehension, while the first column provides results for Year 1 of reform and the second for Year 2 of reform. In each graph, the horizontal axis represents the 2014-15 school performance composite centered at the eligibility threshold. Mean student early literacy and reading comprehension scores, binned by the school's baseline performance composite, appear on the vertical axes, and the eligibility cutoff is indicated by the vertical dashed line. The vertical distance between the fit lines at the cutoff shows the difference in outcomes associated with being in a school assigned to the NCT intervention. The negative effects in Year 1 are apparent in the discontinuity between the lines to the left and right of the cutoff.

Figure 2 ABOUT HERE

In the second year of reform, the discontinuities at the cutoff are smaller and are not consistent across outcomes. We therefore turn to our regression results to interpret both sets of

estimates in Table 3. These estimates represent the local average treatment effect for students in schools near the cutoff. Column 1 shows the estimates within the preferred bandwidth for Year 1 of reform. As the graphical results depict, we find with the RD specification a significant negative effect of NCT on early literacy and reading comprehension in the first year of services. Specifically, student performance on these formative assessments was about 0.2 standard deviations lower in NCT schools than in control schools. An effect of 0.2 standard deviations is considered large in size given the type of educational intervention and research methods used here (Kraft, 2020). These results are robust in terms of significance but vary somewhat in magnitude to alternative bandwidths, shown in Columns 2 and 3. These results meet What Works Clearinghouse standards for integrity of the forcing variable, functional form, and bandwidth. However, our additional robustness check implementing a local randomization estimator does not yield significant results in either year (Appendix A, Table A.2). We therefore highlight the need to interpret these results with caution.

Table 3 ABOUT HERE

Columns 4-6 show that treatment schools rebounded somewhat in the second year of services. Columns 4 shows marginally significant positive effects in Year 2 of about 0.08 to 0.09 standard deviations on early literacy and reading comprehension, respectively. These positive effects are robust to the alternative bandwidths shown in Columns 5 and 6. The positive estimate for reading comprehension conflicts with Figure 2 above because the figure does not adjust for covariates. Both sets of Year 2 estimates were largely robust to the local randomization RD we conducted as an additional robustness check (Appendix A, Table A.2).

We provide results by grade level in Figure 3. In each panel, markers represent effect estimates from separate sharp RD models by grade level and spikes represent 95% confidence

intervals. Reading comprehension and early literacy effects are qualitatively similar across grade levels, with negative coefficient estimates across all grade levels in Year 1 (these effects are significant at $p<0.05$ for all but kindergarten reading comprehension, which was significant at $p<0.10$). In Year 2, the strongest and most consistent positive effects were in reading comprehension in kindergarten. Specifically, kindergarten students in NCT schools performed 0.28 standard deviations higher on the reading comprehension assessment than kindergarten students in control schools. We provide the full results from these models, as well as estimates from RD models within alternative bandwidths, in table form in Appendix B.

Figure 3 ABOUT HERE

Other Student Outcomes

We turn next to the effect of NCT on chronic absenteeism and grade retention, shown in Figure 4. The discontinuity between the two linear splines in Year 1 shows that NCT schools—i.e., those schools to the left of the cutoff—had more chronic absenteeism and grade retention in the first year of the intervention. Table 3 above shows that the effect on grade retention is significant across all bandwidths and the effect on chronic absenteeism is marginally significant in the preferred bandwidth and significant at conventional levels across alternative bandwidths. Specifically, these estimates indicate that grade retention was about 4 percentage points higher in NCT schools in the first year of intervention, while chronic absenteeism was about 3 percentage points higher. These estimates translate to an effect size of 0.19 standard deviations on grade retention and 0.10 standard deviations on chronic absenteeism in Year 1, which are considered moderate effect sizes (Kraft, 2020). These results are largely robust to the local randomization RD shown in Appendix A, Table A.2. We do not detect significant effects on either outcome in the second year of services.

Figure 4 ABOUT HERE

The effects on grade retention were largely concentrated in kindergarten and first grade, as we show in Figure 5 (and in table format in Appendix B). In particular, students in NCT schools were 5.7 percentage points more likely to be retained in kindergarten and 7 percentage points more likely to be retained in first grade in the first year of services, while we do not observe an effect in second grade. In the second year of services, we do not detect an effect on grade retention in kindergarten or second grade, while students in grade 1in NCT schools were 5.8 percentage points more likely to be retained.

The effects on chronic absenteeism were strongest and most consistent in kindergarten. In both the first and second years of the intervention, kindergarten students in NCT schools were approximately 6 percentage points more likely to be chronically absent than their peers in control schools. We do not detect an effect on chronic absenteeism in first or second grade in either year.

Figure 5 ABOUT HERE

Research Question #2: Strategic Reassignment of Teachers to Untested Grades

Two types of strategic staffing could undermine student learning in untested early grades: reassignment of high effectiveness teachers away from these early grades to tested grades, and reassignment of low effectiveness teachers out of tested courses into these early grades. To the extent that these practices occur more in NCT than control schools, they could have driven the negative effects in early literacy and reading comprehension in the first year of services by decreasing K-2 teacher quality in NCT schools. In Table 4, Panel A, Columns 1 and 3 provide the estimated odds ratios of *untested early grade teachers* returning to untested courses, while Columns 2 and 4 provide the estimated odds ratios of *tested teachers* moving to untested courses. Column 2, Row 4 shows that low effectiveness teachers in tested courses across the full sample

were 1.7 times more likely to be reassigned to untested early grades in Year 1 of the intervention—providing evidence that strategic staffing occurred in the first year of NCT across the entire study sample. However, the insignificant coefficient on the interaction term *NCT x low effectiveness* in Column 2, Row 1 suggests that NCT schools did not employ these strategic staffing practices more often than control schools. The empty cells associated with *NCT x high effectiveness* in Row 3, Columns 2 and 4 underscore a salient gap in treatment schools—that there were no high effectiveness teachers of tested grades, as measured by EVAAS on end-of-grade exams, in treatment schools who moved to untested early grades. This finding shows that treatment schools kept 100% of their highly effective teachers who remained in the building assigned to tested grades. By contrast, treatment schools reassigned some of their low and mid effectiveness teachers who remained in the building to untested early grades.

In the second year of services, strategic staffing practices with respect to EVAAS scores followed a less clear pattern. Across the full sample, low effectiveness teachers coming from untested grades were less likely to remain in these untested grades than the reference group of mid-effectiveness teachers (see Column 3, Row 3)—suggesting that the full sample of schools did not engage in strategic staffing to the detriment of younger students by retaining ineffective teachers in early grades. We do not find evidence of differential staffing practices in NCT schools, which would be captured in the interaction terms. Again, in Year 2 of reform, NCT schools had no highly effective teachers in tested grades who were reassigned to untested early grades.

Table 4 ABOUT HERE

Table 4, Panel B shows that across the full sample, novice teachers in untested early grades were less likely than experienced teachers to remain in untested grades the following year

(see Row 3, Columns 1 and 3)—suggesting that on average, treatment and control schools were not disproportionately retaining their inexperienced teachers in untested grades. Again, we do not find evidence of differential practices by treatment condition, although the interaction terms in Column 1 show that NCT schools in the first year of services were descriptively more likely to retain novice teachers in untested courses and to move experienced teachers out of these untested courses. We do not see the same pattern in Year 2 of reform, when early grade student literacy scores rebounded somewhat in NCT schools.

Discussion

This paper provides initial evidence on the effects of school turnaround on early grade student outcomes and strategic staffing in early grades. We find that a school turnaround initiative largely focused on improving instruction in tested grades had modest unintended negative consequences for student learning in younger grades in the first year of reform. Specifically, we find that the NCT initiative increased chronic absenteeism and grade retention and may have produced negative effects on early literacy and reading comprehension in the first year of services. While the negative effects in early literacy and reading comprehension were not robust to the secondary robustness check, the consistent negative estimates across bandwidths using the conventional RD—combined with significantly higher chronic absenteeism and grade retention—may be enough to raise concern about the possibility of unintended consequences of accountability reforms for early learning. The negative effects materialized one year prior to the negative effects that were documented in tested grades in the second year of the intervention (Henry & Harbatkin, 2020). In the second year, we found positive effects on early literacy and reading comprehension, although the magnitude of the increase was smaller than the dip in Year 1.

We highlight that it is possible that scores continued to rebound in subsequent years, ultimately canceling out the negative effects in Year 1 over time. However, a rebound sufficient to yield a net positive effect seems improbable given that the intervention ended and the state withdrew supports from NCT schools. We also note that these effects are not necessarily attributable to the coaching itself, which was largely focused on teachers of tested grades and subjects and which other research suggests teachers perceived in a positive light. Instead, these results may stem from the accountability policy of labeling low performance, simply being designated by the state as a low-performing turnaround school. While prior studies have pointed to the possibility that schools and districts redirect resources away from untested early grades in response to accountability systems (e.g., Chingos & West, 2011; Fuller & Ladd, 2013), this study leverages unique student achievement data from untested early grades to show evidence strongly suggestive of declines on student literacy outcomes in the first year of a turnaround intervention. These findings reinforce the other findings of unintended effects of test-based accountability on student outcomes in untested grades.

One potential mechanism that may help to explain the modest negative effects of NCT on student learning in early grades is that the stigma associated with the turnaround label in NCT schools— combined with demoralization from not receiving additional resources—may have undermined teaching and learning in these untested grades. Any demoralization associated with the stigma may have been exacerbated in early grades because NCT supports were largely directed toward tested grades, where students take tests that count toward school accountability scores. It is possible that the turnaround label reduced morale among teachers in the first year, and that the morale drop was not counteracted by supports for early-grade teachers. The Year 2 reading comprehension increase was largest among kindergarteners—students who would not

have been exposed to the first year of treatment. On the other hand, it is possible that educators were unhappy with the supports they did receive as part of the intervention, leading to low morale and potentially resistance that resulted in underperformance. However, we note that qualitative research suggests that educators who did receive coaching supports were largely satisfied with the services they received (Herman et al., 2019) but not all teachers received coaching and those who did not may have been unhappy to have been left out. Finally, it is possible that the increased coaching in Year 2 produced more positive effects on student outcomes after the initial declines. However, we find this explanation unlikely given that the coaching targeted tested grades and subjects and prior research has shown that the intervention produced negative effects on student achievement in the targeted grades in Year 2 (Henry & Harbatkin, 2020).

In addition to finding negative effects on student outcomes in the RD framework, we also find in our descriptive analysis that schools across the full sample strategically reassigned low effectiveness teachers to untested courses where their students' academic performance would not count toward school accountability scores, though this practice was not more prevalent in NCT than control schools. Because strategic staffing was occurring across our full sample of schools, it does not explain the negative effects on early literacy and reading comprehension in the first year of the intervention in treatment schools. Nevertheless, our findings regarding the strategic assignment of teachers does help to elucidate the types of practices occurring in the schools, which were all designated as low performing by the state. Further, while our analysis of strategic staffing is not causal, it does provide associational evidence that strategic staffing was, in fact, occurring in these schools and potentially to the detriment of younger students. Control schools engaging in these strategic staffing practices is unsurprising; these schools were also designated

as low performing and would therefore be subject to many of the same accountability pressures as turnaround schools.

These findings therefore add some context to the mounting evidence that younger students in low-performing schools may be subject to lower quality teaching than their older peers (Atteberry et al., 2019). To that end, learning loss in early grades may inhibit the sustainability of school turnaround initiatives, which have two central aims—to rapidly improve student performance and then sustain those improvements over multiple years (Aladjem et al., 2010; Herman et al., 2008). As schools with limited human resources prioritize rapid improvement in tested grades, they may in turn undermine longer term sustainability of a turnaround. By strategically reassigning low-effectiveness teachers to untested courses, low-performing schools are not only redirecting critical resources in the form of teacher quality away from early grades, but also destabilizing school turnaround processes across all grade levels. Given that low-performing schools experience intense pressure to improve student outcomes under high-stakes accountability systems, future research should investigate whether there are avenues for accountability-driven turnaround that do not reduce resources for students in early grades.

A limitation of this study, given our focus solely on untested early grades, is that our sample has limited power to detect effects within the RD design. The small sample may have additionally limited our ability to detect significant differences in strategic staffing practices between NCT and control schools—in particular with regards to high effectiveness teachers in untested early grades, who were descriptively less likely to return to untested subjects in treatment than control schools.

Conclusion

This study provides information for stakeholders, including policymakers, parents, and educators, who are interested in early childhood investments and their subsequent effects on student outcomes. We find that the NCT initiative increased chronic absenteeism and grade retention in the first year of the reform by a small degree and had null effects in the second year. Also in the first year of the intervention, we find suggestive evidence of negative effects on early literacy and reading comprehension with scores rebounding partially in the second reform year. In our descriptive staffing analysis, we find that across the entire sample of low-performing schools, schools strategically reassigned low effectiveness teachers from tested to untested courses, potentially weakening low performing schools' performance on accountability exams when these students progress into later grades. Further research is needed to better understand the impact of school turnaround on early-grade student outcomes and to explore possible mechanisms that could alleviate accountability pressures on schools to engage in strategic staffing. In some settings, offering financial incentives for recruiting and retaining effective teachers and principals has helped turnaround schools to improve student achievement and sustain improvements over time (Henry et al., 2020). Even in schools that successfully achieve rapid gains in their lowest performing schools, state and district monitoring ought to focus some attention on early grade outcomes—including hiring and placement of effective teachers—in order to better position these schools for sustained improvements.

Finally, research on longer term effects of pre-K and other early interventions may need to examine strategic staffing as a possible explanation for the fade-out and even reversal of effects. Lower quality teachers in early grades may not be able to amplify the skills of higher performing students, may teach more basic skills, and may lack the skills to effectively differentiate instruction. Negative effects on younger students may be magnified if children

participating in targeted pre-K programs attend lower performing schools that are subject to test-based accountability pressures.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135. https://doi.org/10.1086/508733

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54–76. https://doi.org/10.3102/0162373716663646

Aladjem, D. K., Birman, B. F., Orland, M., Harr-Robins, J., Heredia, A., Parrish, T. B., & Ruffini, S. J. (2010). *Achieving dramatic school improvement: An exploratory study*. U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service. https://eric.ed.gov/?id=ED526783

Amplify Education. (2014). *Validation data and reports*. https://www.ode.state.or.us/wma/teachlearn/testing/resources/mclass_reading_3d_validation.pdf

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, *131*(3), 1415–1453. https://doi.org/10.1093/qje/qjw016

Atteberry, A., Bassok, D., & Wong, V. C. (2019). The effects of full-day prekindergarten: Experimental evidence of impacts on children's school readiness. *Educational Evaluation and Policy Analysis*, *41*(4), 537–562. https://doi.org/10.3102/0162373719872197

Ballou, D., & Springer, M. G. (2016). Has NCLB encouraged educational triage? Accountability and the distribution of achievement gains. *Education Finance and Policy*, *12*(1), 77–106. https://doi.org/10.1162/EDFP_a_00189

Bassok, D. (2010). Do Black and Hispanic children benefit more from preschool? Understanding differences in preschool effects across racial groups. *Child Development*, *81*(6), 1828–1845. https://doi.org/10.1111/j.1467-8624.2010.01513.x

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326. https://doi.org/10.3982/ECTA11757

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Carlson, D., & Lavertu, S. (2018). School improvement grants in Ohio: Effects on student achievement and school administration. *Educational Evaluation and Policy Analysis*, *40*(3), 287–315. https://doi.org/10.3102/0162373718760218

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. Senate. *Journal of Causal Inference*, *3*(1), 1–24. https://doi.org/10.1515/jci-2013-0010

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2016). Inference in regression discontinuity designs under local randomization. *The Stata Journal: Promoting Communications on Statistics and Stata*, *16*(2), 331–367. https://doi.org/10.1177/1536867X1601600205

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, *126*(4), 1593–1660. https://doi.org/10.1093/qje/qjr041

Chingos, M. M., & West, M. R. (2011). Promotion and reassignment in public school districts: How do schools respond to differences in teacher effectiveness? *Economics of Education Review*, *30*(3), 419–433. https://doi.org/10.1016/j.econedurev.2010.12.011

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778–820. https://doi.org/10.3368/jhr.XLI.4.778

Coburn, C. E., & Woulfin, S. L. (2012). Reading coaches and the relationship between policy and practice. *Reading Research Quarterly*, *47*(1), 5–30. https://doi.org/10.1002/RRQ.008

Cohen-Vogel, L. (2011). "Staffing to the test": Are today's school personnel practices evidence based? *Educational Evaluation and Policy Analysis*, *33*(4), 215–229. https://doi.org/10.3102/0162373711419845

Currie, J. (2001). early childhood education programs. *Journal of Economic Perspectives*, *15*(2), 213–238. https://doi.org/10.1257/jep.15.2.213

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*(3), 111–134. https://doi.org/10.1257/app.1.3.111

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. https://doi.org/10.1177/0895904817719520

Dynarski, S., Hyman, J., & Schanzenbach, D. W. (2013). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. *Journal of Policy Analysis and Management*, *32*(4), 692–717. https://doi.org/10.1002/pam

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons From Chicago's low-performing schools. *American Educational Research Journal*, *44*(3), 594–630. https://doi.org/10.3102/0002831207306767

Fuller, S. C., & Ladd, H. F. (2013). School-based accountability and the distribution of teacher quality across grades in elementary school. *Education Finance and Policy*, *8*(4), 528–559. https://doi.org/10.1162/EDFP_a_00112

Gamoran, A. (1987). The stratification of high school learning opportunities. *Sociology of Education*, *60*(3), 135–155. https://doi.org/10.2307/2112271

Glennie, E. J., Mason, M., & Edmunds, J. A. (2016). Retention and satisfaction of novice teachers: Lessons from a school reform model. *Journal of Education and Training Studies*, *4*(4), 244–258. https://doi.org/10.11114/jets.v4i4.1458

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96–104. https://doi.org/10.3102/0013189X15575031

Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Institute for the Development of Educational Achievement.

Gottfried, M. A., & Hutt, E. L. (Eds.). (2019). *Absent from school: Understanding and addressing student absenteeism.* Harvard Education Press.

Graham, L. J., White, S. L. J., Cologon, K., & Pianta, R. C. (2020). Do teachers' years of experience make a difference in the quality of teaching? *Teaching and Teacher Education*, *96*, 1–10. https://doi.org/10.1016/j.tate.2020.103190

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). The micropolitics of educational inequality: The case of teacher–student assignments. *Peabody Journal of Education*, *90*(5), 601–614. https://doi.org/10.1080/0161956X.2015.1087768

Grissom, J. A., Kalogrides, D., & Loeb, S. (2017). Strategic staffing? How performance pressures affect the distribution of teachers within schools and resulting student achievement. *American Educational Research Journal*, *54*(6), 1079–1116. https://doi.org/10.3102/0002831217716301

Hamilton, M. P., Heilig, J. V., & Pazey, B. L. (2014). A nostrum of school reform? Turning around reconstituted urban Texas high schools. *Urban Education*, *49*(2), 182–215. https://doi.org/10.1177/0042085913475636

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*. https://doi.org/10.1080/19345747.2020.1814464

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. *Educational Evaluation and Policy Analysis*, *42*(2), 232–256. https://doi.org/10.3102/0162373720908600

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., & Redding, S. (2008). *Turning around chronically low-performing schools: A practice guide* (NCEE 2008-4020; IES Practice Guide). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/wwc/PracticeGuide.aspx?sid=7

Herman, R., Johnston, W. R., Migacheva, K., & Tosh, K. (2019). *delivery of educational support services to low-performing schools in North Carolina*. https://stateboard.ncpublicschools.gov/resources/other-reports/36001servicedeliverybriefbrochure.pdf

Hess, G. A. (2003). Reconstitution—three years later: Monitoring the effect of sanctions on Chicago high schools. *Education and Urban Society*, *35*(3), 300–327. https://doi.org/10.1177/0013124503035003004

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*(6), 351–360. https://doi.org/10.3102/0013189X08323842

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *The Quarterly Journal of Economics*, *131*(1), 157–218. https://doi.org/10.1093/qje/qjv036

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *118*(3), 843–877. https://doi.org/10.1162/00335530360698441

Johnson, R. C., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy*, *11*(4), 310–349. https://doi.org/10.1257/pol.20180510

Jordan, P. W., & Miller, R. (2017). *Who's in: Chronic absenteeism under the Every Student Succeeds Act.* FutureEd at Georgetown University.

Kalogrides, D., Loeb, S., & Béteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*(2), 103–123.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* MET Project. Bill & Melinda Gates Foundation.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Kraft, M. A., Papay, J. P., & Chi, O. L. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management*, *39*(2), 315–347. https://doi.org/10.1002/pam.22193

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, *29*(3), 426–450. https://doi.org/10.1002/pam.20504

Ladd, H. F., & Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy*, *12*(2), 241–279. https://doi.org/10.1162/EDFP_a_00194

Li, W., Duncan, G. J., Magnuson, K., Schindler, H. S., Yoshikawa, H., & Leak, J. (2020). Timing in early childhood education: How cognitive and achievement program impacts

vary by starting age, program duration, and time since the end of the program. Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai20-201

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly*, *45*, 155–176. https://doi.org/10.1016/j.ecresq.2018.03.005

Mathis, W. J. (2009). NCLB's ultimate restructuring alternatives: Do they improve the quality of education? In *Education Policy Research Unit*. Education Policy Research Unit. https://eric.ed.gov/?id=ED507355

Maxcy, B. (2009). New public management and district reform: Managerialism and deflection of local leadership in a Texas school district. *Urban Education*, *44*(5), 489–521. https://doi.org/10.1177/0042085908318778

Meyers, C. V., & Hitt, D. H. (2018). Planning for school turnaround in the United States: An analysis of the quality of principal-developed quick wins. *School Effectiveness and School Improvement*, *29*(3), 362–382. https://doi.org/10.1080/09243453.2018.1428202

Meyers, C. V., & VanGronigen, B. A. (2018). So many educational service providers, so little evidence. *American Journal of Education*, *125*(1), 109–139. https://doi.org/10.1086/699823

Murillo, E. G., & Flores, S. Y. (2002). Reform by shame: Managing the stigma of labels in high stakes testing. *The Journal of Educational Foundations*, *16*(2), 93–108.

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

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., Duncan, G. J., Dynarski, M., Magnuson, K. A., & Weiland, C. (2017). *The current state of scientific knowledge on pre-kindergarten effects*. Brookings Institution and the Duke Center for Child and Family Policy. https://www.fcd-us.org/current-state-scientific-knowledge-pre-kindergarten-effects/

Redding, C., & Nguyen, T. D. (2020). The relationship between school turnaround and student outcomes: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(4), 0162373720949513. https://doi.org/10.3102/0162373720949513

Rice, J. K., & Malen, B. (2003). The human costs of education reform: The case of school reconstitution. *Educational Administration Quarterly*, *39*(5), 635–666. https://doi.org/10.1177/0013161X03257298

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

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,* online. https://doi.org/10.3102/00028312211060855

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, S. W., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool Study through age 40*. High/Scope Press.

Smith, K. C., Amendum, S. J., & Jang, B. G. (2020). Predicting performance on a 3rd grade high-stakes reading assessment. *Reading & Writing Quarterly*, *36*(4), 365–378. https://doi.org/10.1080/10573569.2019.1649612

Strunk, K. O., Marsh, J. A., Bush-Mecenas, S. C., & Duque, M. R. (2016). The best laid plans: An examination of school plan quality and implementation in a school improvement initiative. *Educational Administration Quarterly*, *52*(2), 259–309. https://doi.org/10.1177/0013161X15616864

Strunk, K. O., Marsh, J. A., Hashim, A. K., Bush-Mecenas, S., & Weinstein, T. (2016). The impact of turnaround reform on student outcomes: Evidence and insights from the Los Angeles Unified School District. *Education Finance and Policy*, *11*(3), 251–282. https://doi.org/10.1162/EDFP

VanGronigen, B. A., & Meyers, C. V. (2019). How state education agencies are administering school turnaround efforts: 15 years after No Child Left Behind. *Educational Policy*, *33*(3), 423–452. https://doi.org/10.1177/0895904817691846

Wallace Foundation. (2010). *The school turnaround field guide*. https://www.wallacefoundation.org/knowledge-center/pages/the-school-turnaround-field-guide.aspx

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*(6), 2112–2130. https://doi.org/10.1111/cdev.12099

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook, version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

Woulfin, S. L. (2015). Catalysts of change: An examination of coaches' leadership practices in framing a reading reform. *Journal of School Leadership*, *25*(3), 526–557. https://doi.org/10.1177/105268461502500309

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). *SAS EVAAS statistical models*. SAS Institute.

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. https://doi.org/10.3102/0162373717705729

Tables

Table 1. Baseline school sample characteristics conditional on forcing variable

	Treatment	Control	*p*-value
Student demographics			
Economically disadvantaged percent	88.70	89.23	0.919
Black percent	67.21	50.52	0.365
Hispanic percent	13.98	25.32	0.271
Per pupil spending	9,539.01	9,694.70	0.873
Average daily membership	412.99	413.77	0.994
Teacher demographics			
Novice teacher rate	37.73	46.25	0.252
Fully licensed teacher rate	93.94	95.16	0.765
School performance			
School growth	-3.55	-3.25	0.935
N schools	38	137	

Estimates from regression discontinuity (RD) design with covariate listed in row as outcome and triangular kernel. All analyses are of school-level means because treatment assignment occurred at the school level. School growth is the school-level value of EVAAS, the state's value-added measure, which has a mean of zero and can theoretically range from negative to positive infinity. Most schools in the state fall between -2 and +2, which the state classifies as meeting expected growth. Schools with an EVAAS score above two are classified as exceeding expected growth. Schools with an EVAAS score below two—as both treatment and control schools have in this sample of low-performing schools—are classified as failing to meet expected growth.

Table 2. Descriptive statistics for student outcome variables

Panel A. Year 1 of intervention

	Full sample		Within bandwidth	
	Treatment	Control	Treatment	Control
Early literacy	-0.078	0.065	0.008	0.021
	(1.005)	(0.985)	(0.999)	(0.973)
N students	*6345*	*22941*	*2346*	*1755*
Reading comprehension	-0.116	0.077	-0.025	0.030
	(1.008)	(0.984)	(1.014)	(0.975)
N students	*6514*	*22619*	*2650*	*1735*
Chronic absenteeism	0.104	0.080	0.096	0.076
	(0.306)	(0.271)	(0.295)	(0.266)
N students	*7981*	*26860*	*3081*	*2018*
Grade retention	0.055	0.044	0.053	0.039
	(0.229)	(0.204)	(0.223)	(0.193)
N students	*7981*	*26860*	*3081*	*2018*

Panel B. Year 2 of intervention

	Full sample		Within bandwidth	
	Treatment	Control	Treatment	Control
Early literacy	-0.135	0.077	-0.070	-0.097
	(1.009)	(0.979)	(0.997)	(0.984)
N students	*5921*	*22071*	*2361*	*1604*
Reading comprehension	-0.112	0.087	0.005	-0.058
	(1.020)	(0.979)	(1.033)	(0.948)
N students	*5555*	*20799*	*2302*	*1520*
Chronic absenteeism	0.145	0.113	0.143	0.104
	(0.352)	(0.317)	(0.350)	(0.305)
N students	*7626*	*26384*	*2884*	*1927*
Grade retention	0.054	0.044	0.044	0.044
	(0.225)	(0.205)	(0.205)	(0.204)
N students	*7626*	*26384*	*2884*	*1927*

Means are presented with standard deviations in parentheses. Full sample includes students in all treatment or control schools. Within bandwidth includes students in treatment or control schools that are within the optimal bandwidth for our regression discontinuity models. This bandwidth is calculated using the bandwidth selection procedure described by Calonico, Cattaneo, & Titiunik (2014).

Table 3. Estimates on grades K-2 early literacy, reading comprehension, chronic absenteeism, & grade retention

	Year 1			Year 2		
	(1) CCT	(2) 150% CCT	(3) 200% CCT	(4) CCT	(5) 150% CCT	(6) 200% CCT
Early literacy	-0.222***	-0.124***	-0.072*	0.079+	0.098**	0.106***
	(0.0454)	(0.0367)	(0.0312)	(0.0467)	(0.0377)	(0.0319)
N	*29286*	*29286*	*29286*	*27992*	*27992*	*27992*
N students within bandwidth	*4101*	*6520*	*9348*	*3965*	*6148*	*8793*
Reading comprehension	-0.232***	-0.100**	-0.059+	0.086+	0.171***	0.128***
	(0.0468)	(0.0384)	(0.0328)	(0.0510)	(0.0414)	(0.0347)
N	*29133*	*29133*	*29133*	*26354*	*26354*	*26354*
N students within bandwidth	*4385*	*6790*	*9463*	*3822*	*5874*	*8440*
Chronic absenteeism	0.029+	0.034**	0.026*	0.007	0.012	0.009
	(0.0161)	(0.0129)	(0.0110)	(0.0179)	(0.0152)	(0.0133)
N	*34841*	*34841*	*34841*	*34010*	*34010*	*34010*
N students within bandwidth	*5099*	*7951*	*11376*	*4811*	*7576*	*10999*
Grade retention	0.040**	0.033**	0.022*	-0.001	-0.001	0.003
	(0.0129)	(0.0106)	(0.0090)	(0.0138)	(0.0109)	(0.0090)
N	*34841*	*34841*	*34841*	*34010*	*34010*	*34010*
N students within bandwidth	*5099*	*7951*	*11376*	*4811*	*7576*	*10999*
Bandwidth	2.9	4.3	5.7	2.9	4.3	5.7
N schools below cutoff	14	22	27	14	22	27
N schools above cutoff	12	19	29	12	19	29

Estimates from sharp RD using triangular kernel, linear splines, and heteroskedasticity-robust standard errors. CCT refers to the RD bandwidth selection procedure described by Calonico, Cattaneo, & Titiunik (2014). Early literacy and reading comprehension models are conditioned on beginning-of-year scores, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, and days between beginning and end of year assessments. All models control for school and student covariates. School covariates include minority percentage, economically disadvantaged percentage, per pupil expenditures (PPE) and PPE squared, and enrollment and enrollment squared. Student covariates include grade level with kindergarten as the reference category, gender, race/ethnicity with white as the reference category, disabled, limited English proficient, over-age for grade, and nonstructural transfer in.
$^{+}p<.10$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Table 4. Logistic regression estimates of strategic staffing by teacher effectiveness score and experience across treatment and control schools

Panel A. Teacher effectiveness score

Odds ratio of teaching in untested grade in year *t*+1	Year 1 (2016)		Year 2 (2017)	
Teaching assignment in year t →	*Untested early grades in 2015*	*Tested grades/subjects in 2015*	*Untested early grades in 2016*	*Tested grades/subjects in 2016*
	(1)	(2)	(3)	(4)
NCT x low effectiveness	0.818 [-0.847]	0.768 [-0.583]	0.944 [-0.193]	1.184 [0.291]
NCT x mid effectiveness	1.192 [0.834]	0.648 [-1.430]	0.775 [-1.292]	0.807 [-0.527]
NCT x high effectiveness	0.631 [-1.344]	--[a]	0.725 [-1.005]	--[a]
Low effectiveness	0.889 [-0.785]	1.702* [2.053]	0.668** [-2.618]	1.257 [0.648]
High effectiveness	1.144 [0.846]	0.200 [-1.595]	0.958 [-0.262]	0.320+ [-1.674]
Constant	1.634 [0.201]	0.000* [-2.065]	47.146 [1.569]	0.000* [-2.456]
N	1808	1466	1706	1421

Panel B. Teacher experience

Odds ratio of teaching in untested course in year *t*+1	Year 1 (2016)		Year 2 (2017)	
Teaching assignment in year t →	*Untested early grades in 2015*	*Tested grades/subjects in 2015*	*Untested early grades in 2016*	*Tested grades/subjects in 2016*
	(1)	(2)	(3)	(4)
NCT x novice	1.400	0.614	0.844	1.600
	[1.360]	[-1.213]	[-0.792]	[0.857]
NCT x experienced	0.793	0.851	0.794	0.601
	[-1.304]	[-0.506]	[-1.154]	[-1.159]
Novice	0.505***	1.220	0.668**	0.719
	[-4.835]	[0.821]	[-3.280]	[-1.042]
Constant	4.344	0.000+	34.082	0.000*
	[0.640]	[-1.915]	[1.390]	[-2.382]
N	1903	1773	1789	1737

Estimates from logistic regressions and reported as odds ratios. T-statistics are reported in brackets. Robust standard errors clustered at the school level. Low effectiveness is defined as an EVAAS score of less than -2, which the state categorizes as not meeting expected growth. Mid effectiveness is defined as an EVAAS score between -2 and 2, which the state categorizes as meeting expected growth. High effectiveness is defined as an EVAAS score greater than 2, which the state categorizes as exceeding expected growth. Novice is defined as fewer than 4 years of experience. Teacher covariates include gender and race/ethnicity with white as the reference category. School covariates include minority percentage, economically disadvantaged percentage, per pupil expenditures (PPE) and PPE squared, and enrollment and enrollment squared. $^{+}p<0.10$ $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

[a] NCT x high effectiveness is omitted because no teachers of tested grades or subjects who taught in schools assigned to NCT and were rated as highly effective in year *t* returned to the same school and taught in untested grades in year *t*+1

Figures

Figure 1. North Carolina Transformation Theory of Change

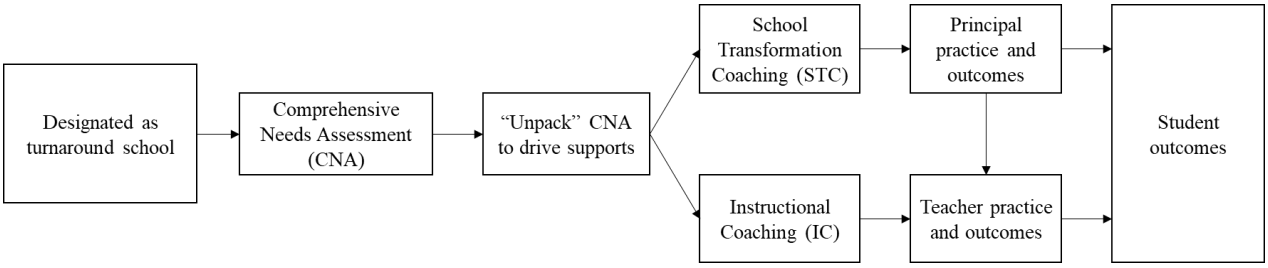
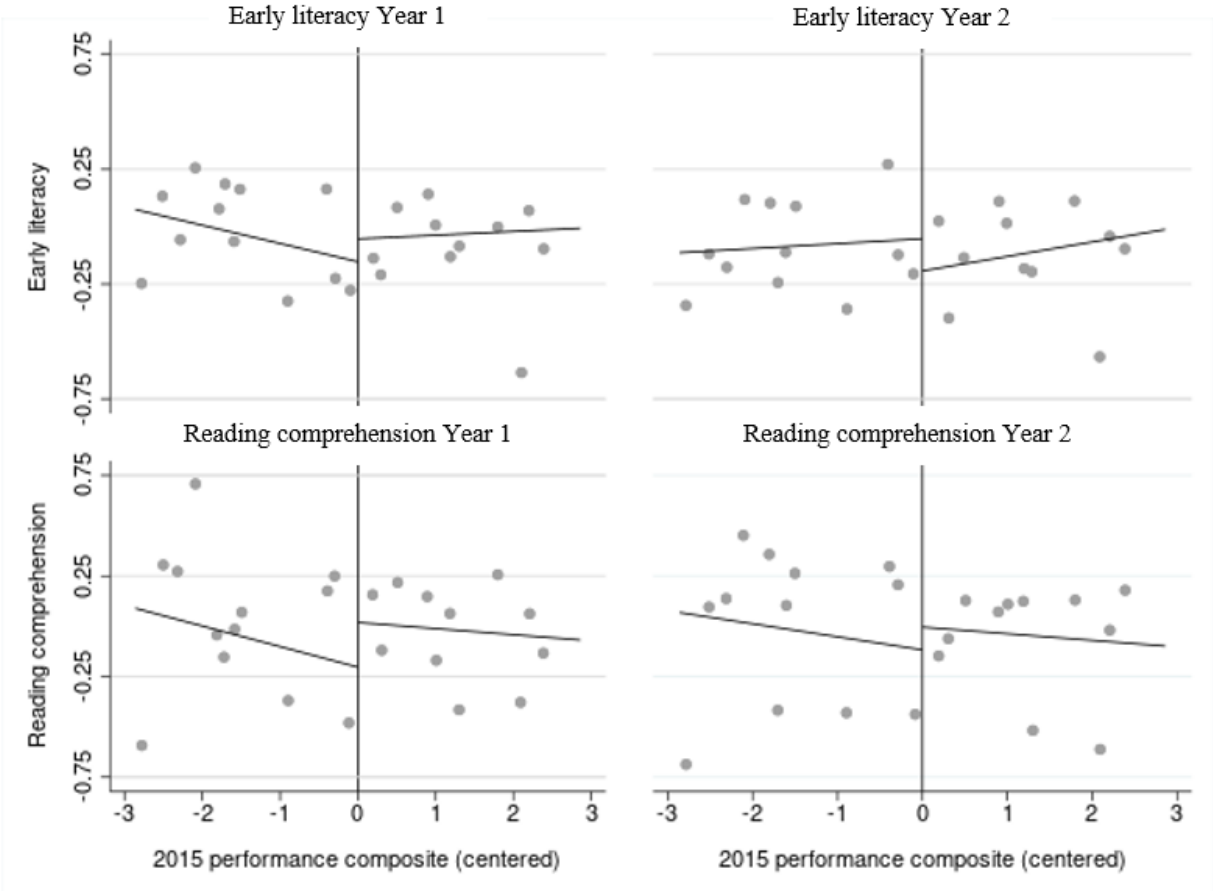
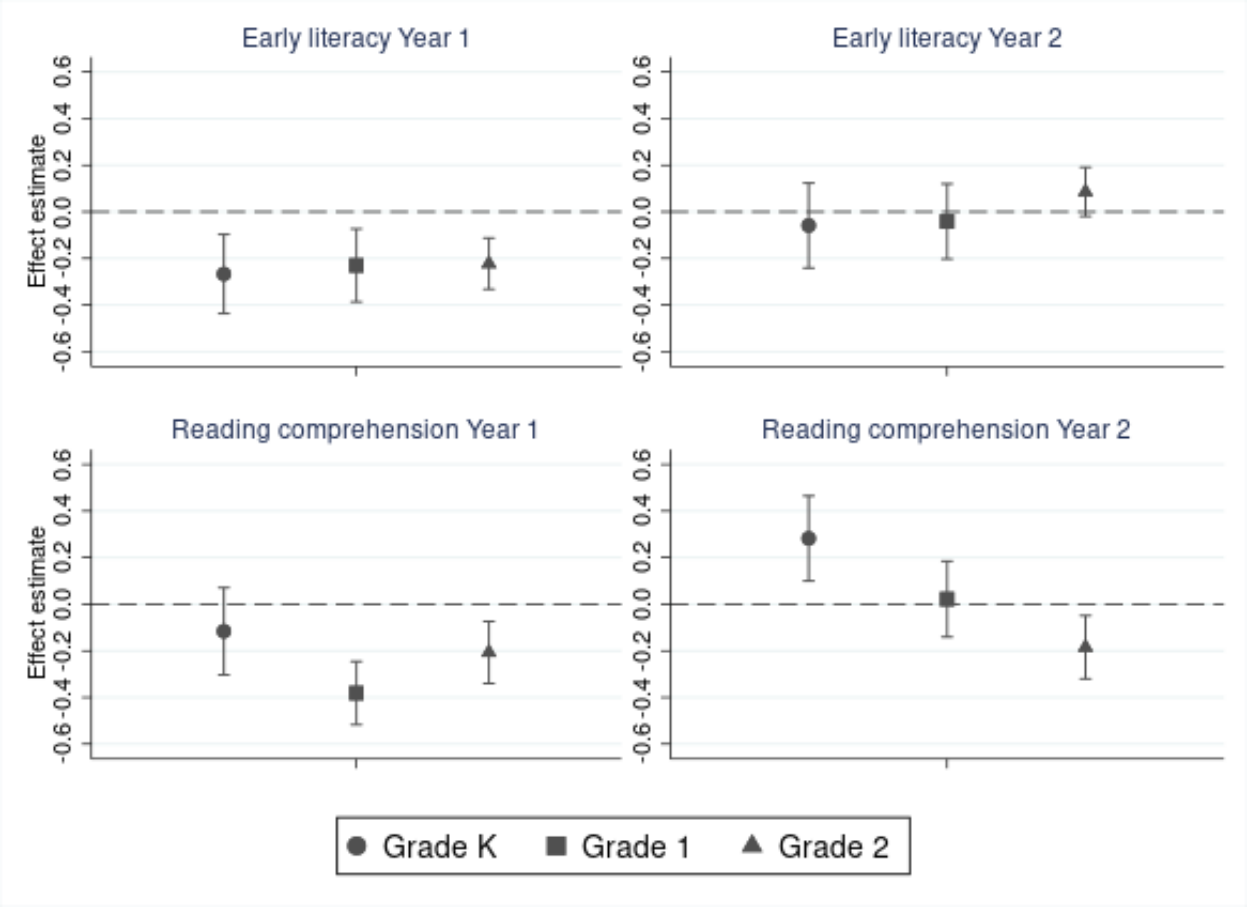


Figure 2. The effects of NCT on early literacy and reading comprehension



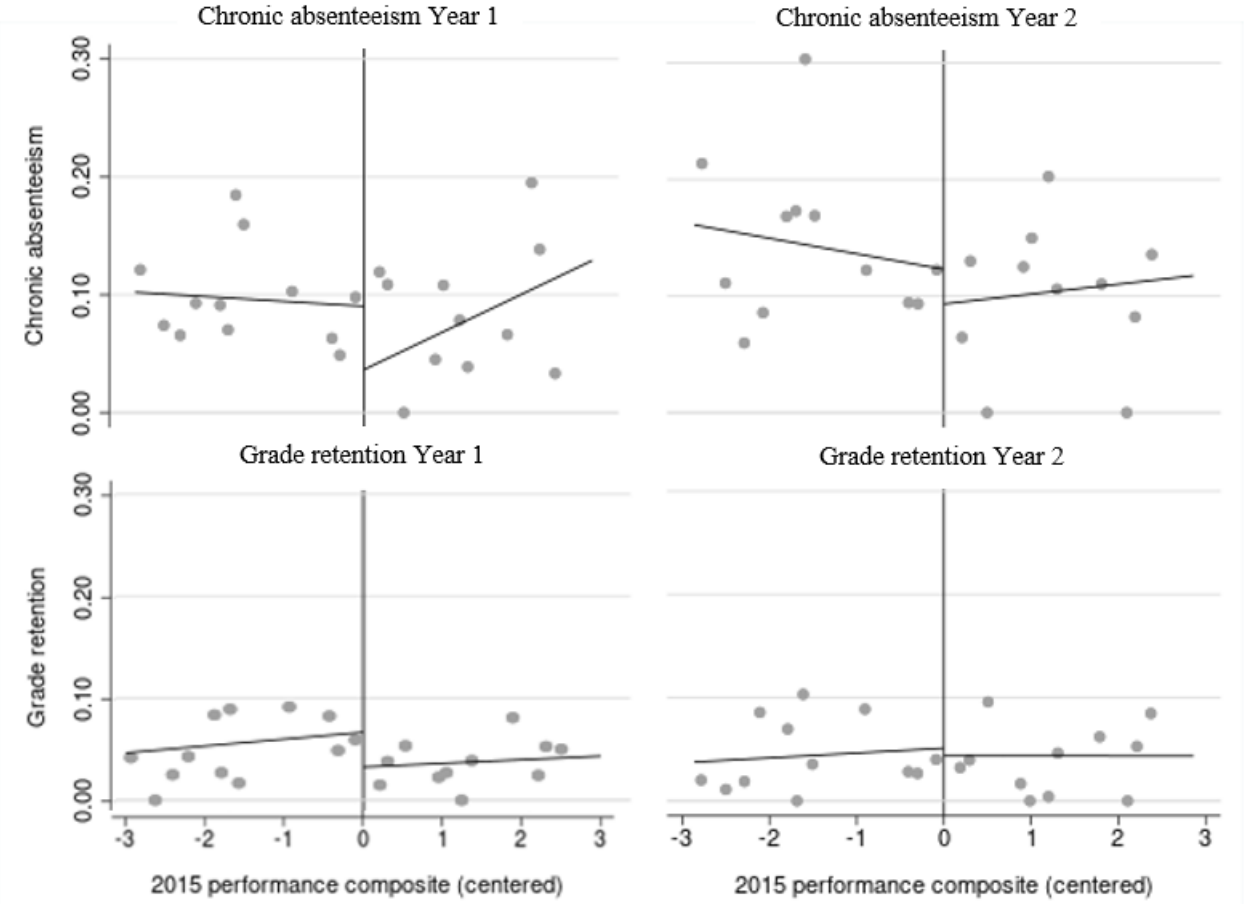
NOTE: Markers represent bin averages of school-level means. Bins are groups of schools with similar baseline proficiency rates. Line is linear fit.

Figure 3. Effect estimates of NCT on early literacy and reading comprehension by grade



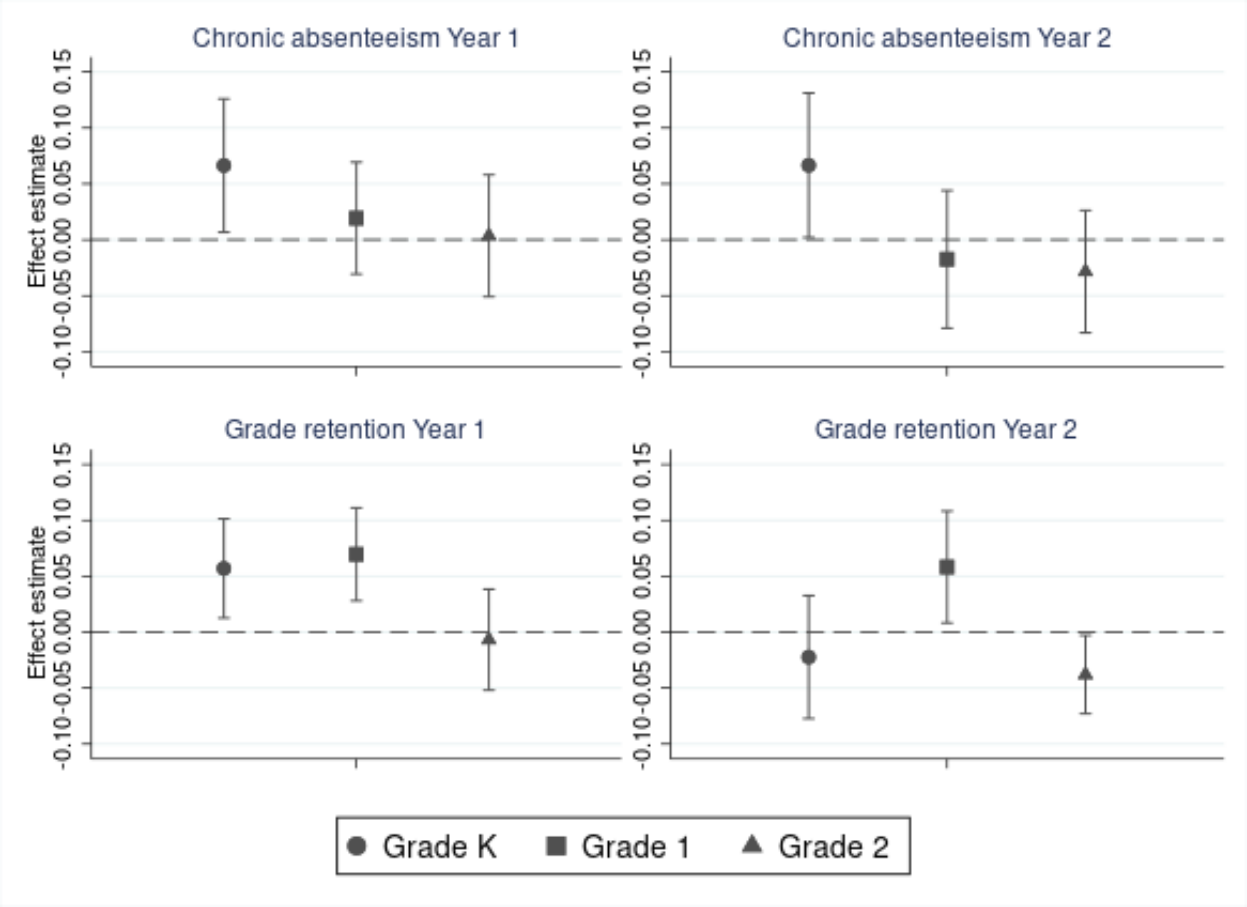
NOTE: Estimates from sharp RD within the preferred CCT bandwidth and using triangular kernel, linear splines, and heteroskedasticity-robust standard errors. Markers represent effect estimates and spikes represent 95% confidence intervals.

Figure 4. The effects of NCT on chronic absenteeism and grade retention



NOTE: Schools were divided into groups or "bins" based on their 2015 performance composite. Then, for each bin, the average of the school-level mean outcome (chronic absenteeism or grade retention) was calculated and plotted as a dot on the figure. Line is linear fit.

Figure 5. Effect estimates of NCT on chronic absenteeism and grade retention by grade



NOTE: Estimates from sharp RD within the preferred CCT bandwidth and using triangular kernel, linear splines, and heteroskedasticity-robust standard errors. Markers represent effect estimates and spikes represent 95% confidence intervals.

Appendix A

In this appendix, we describe four core assumptions of the RD design and then provide evidence that the data in this study meet those assumptions. The first assumption to the validity of the RD design is that there should be no manipulation of the forcing variable. Because the state of North Carolina determined the cutoff score on the forcing variable after schools administered end-of-year exams, manipulation of the forcing variable by schools is highly unlikely. Nevertheless, below we demonstrate both the graphical and statistical integrity of the forcing variable. Specifically, Figure A.1 shows the density of the forcing variable across all eligible schools. The dashed vertical line at zero represents the cutoff score. The lack of a difference in density around the cutoff score demonstrates that there was no manipulation of the forcing variable. We also conducted a McCrary test to test the assumption of no manipulation. The test fails to reject the null of continuity of the density of the forcing variable ($p=.6510$), providing further evidence that the value of the school performance composite was not manipulated to influence treatment assignment near the cutoff.

The second assumption to the validity of the RD design is that the functional form of the relationship between the outcome and forcing variable is correctly specified on both sides of the cutoff value. We estimate separate local linear regressions on either side of the cutoff to meet this condition. Figures 2 and 3, included in the main text, visually demonstrate that the relationships between the outcome variables and forcing variable are linear.

The third assumption for the consistency of the sharp RD estimates is that the relationship between the forcing variable and outcome should be consistent in the absence of the intervention. This assumption cannot be tested directly because we cannot observe outcomes for treatment schools in the absence of treatment. Nevertheless, below we provide two indirect tests of the

continuity of the outcome-forcing variable. First, we test the baseline equivalence of key covariates related to student reading scores across the treatment and control samples, conditional on the forcing variable. As shown in Table 1 of the main text, the p-values associated with the key school-level student demographics, teacher demographics, and school performance covariates are all insignificant, suggesting that our treatment and control samples are balanced on observable characteristics and that the assumption of continuity of the outcome-forcing variable in the absence of treatment likely holds. Second, we graphically examine the relationship between the outcomes and the forcing variable across the full sample. Appendix Figures A.3 and A.4 show no evidence of a discontinuity in the relationship away from the cutoff.

Lastly, the fourth assumption of the sharp RD is that there is no differential attrition across the treatment and control samples. Across first and second year of the intervention, two schools in the control sample closed. As shown in Table A.1, we estimated overall and differential levels of attrition at the school level using a sharp RD and controlling for the forcing variable. We find that the overall and differential levels of attrition are considered low based on the cautious boundary established by the What Works Clearinghouse (2020).

Due to the limited number of schools within the optimal bandwidth, we also estimate the effect of NCT using a local randomization RD design (Cattaneo et al., 2015) as an additional validity check. The local randomization RD design relies on the assumption that treatment is randomly assigned in a small window around the cutoff where covariates are very well balanced. Under this assumption, estimation and inference can be pursued using randomization methods. We use the *rdlocrand* package in Stata to estimate windows near the cutoff where the assumption of randomized treatment assignment is most plausible and to estimate the local randomization

RD models (Cattaneo et al., 2016). The local randomization RD estimates are displayed in Table A.2.

In the first year of services (see Panel A), we consistently find null effect estimates of NCT on early literacy and reading comprehension. These findings are contrary to the estimates from the sharp RD models, which found negative effects on early literacy and reading comprehension in the first year. As such, we view our sharp RD models as providing suggestive evidence of negative effects on student literacy outcomes in Year 1. Consistent with the sharp RD results, we do find evidence from the local randomization RD that rates of chronic absenteeism and grade retention increased in the first year of reform. These results are robust across most window lengths.

In the second year of services (see Panel B of Table A.2), we find consistently positive effects on early literacy and reading comprehension, though the statistical significance of these effects varies across windows. These results support the findings of our sharp RD models that literacy outcomes rebounded in the second year of reform. We also find that chronic absenteeism increased in Year 2 of reform across all windows. Lastly, consistent with the sharp RD results, we do not find significant effects on grade retention in the second year of services.

References

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, *3*(1), 1-24. https://doi.org/10.1515/jci-2013-0010

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2016). Inference in regression discontinuity designs under local randomization. *Stata Journal*, 16, 331–367.

Table A.1 Attrition at the school level

	Year 1 & Year 2
β_{treat}	0.000
$\beta_{control}$	0.021
$\beta_{overall}$	0.010
β_{diff}	-0.021
(SE)	(0.023)

Estimates from sharp RD predicting attrition at the school level and controlling for the forcing variable with a triangular kernel.

Table A.2 Local randomization RD estimates on early literacy, reading comprehension, chronic absenteeism, & grade retention

Panel A. Year 1

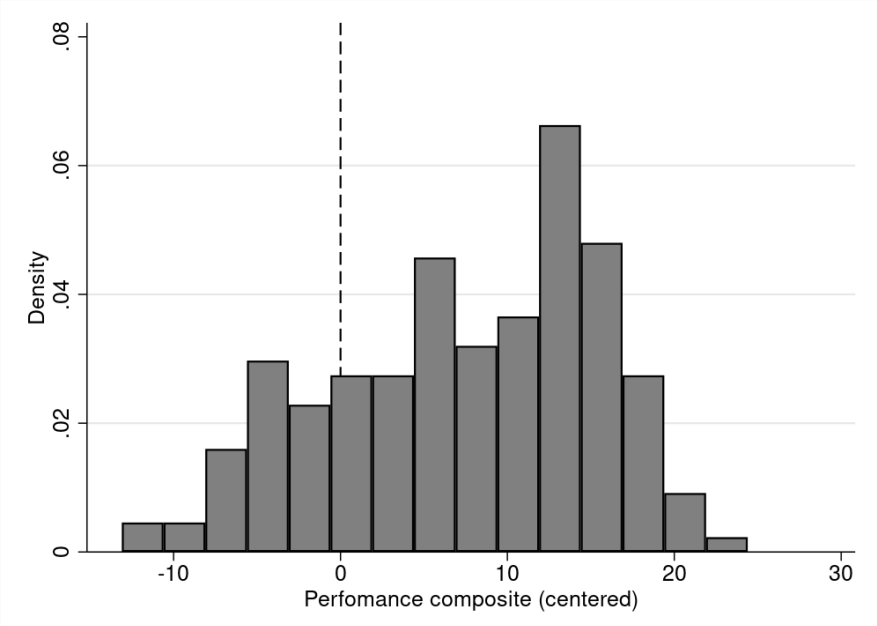
Window length	2.3	2.4	2.5	2.8	2.9	3.2
Early literacy	0.003	0.005	0.017	-0.003	0.041	0.041
Reading comprehension	-0.014	-0.010	0.016	-0.035	0.021	0.021
Chronic absenteeism	0.018*	0.019*	0.018*	0.020*	0.011	0.011
Grade retention	0.020**	0.019**	0.015*	0.014*	0.008	0.008

Panel B. Year 2

Window length	2.3	2.4	2.5	2.8	2.9	3.2
Early literacy	0.059	0.059*	0.054	0.030	0.064*	0.064*
Reading comprehension	0.091**	0.084**	0.088**	0.025	0.063*	0.063*
Chronic absenteeism	0.035**	0.034**	0.032**	0.039***	0.026**	0.026**
Grade retention	0.007	0.005	0.003	0.000	-0.001	-0.001

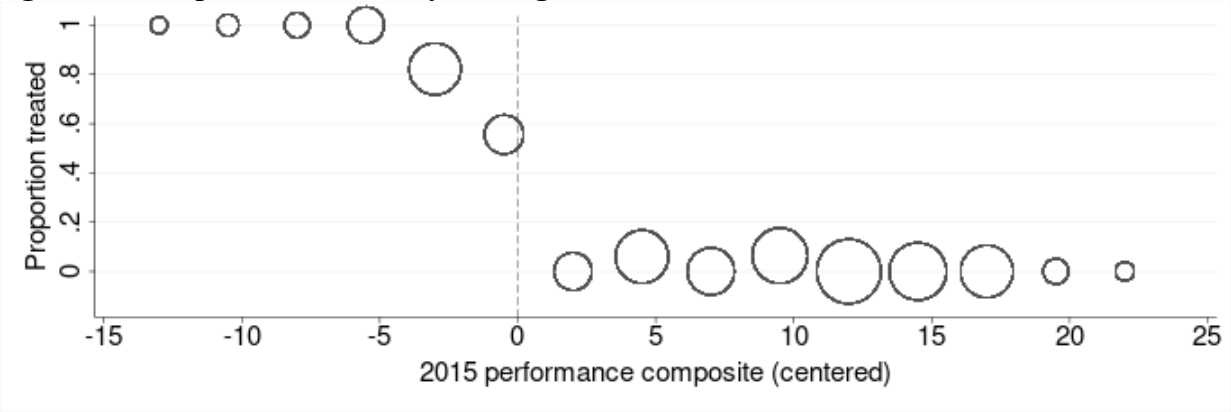
NOTE: Window length represents length on either side of the cutoff. For example, 2.3 runs from -2.3 to +2.3. Estimates from local randomization RD using uniform kernel. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A.1 Graphical integrity of the forcing variable



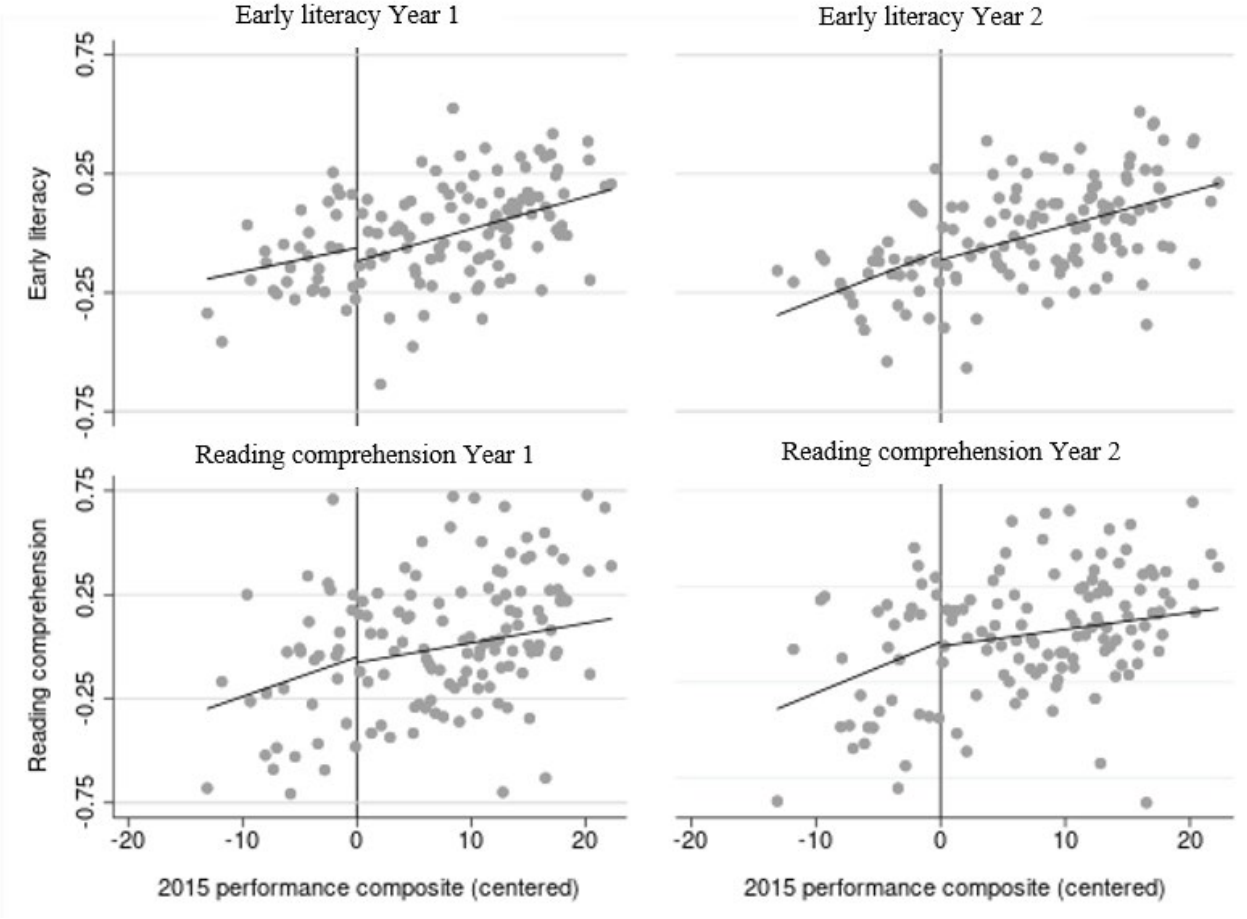
NOTE: The cutoff score for NCT participation was 31.1 for schools enrolling K-2 students, with the 38 schools scoring below 31.1 being targeted for services. There was a different eligibility cutoff for elementary, middle, and high schools. 31.1 was the eligibility cutoff for elementary schools. The state classified schools with a terminal grade of 6 or below as elementary, and as 7 or 8 as middle. Two K-8 schools were therefore classified as middle and subject to the middle school eligibility threshold of 33.8. We centered all schools at 0 according to the appropriate eligibility threshold given their terminal grade level. Bin width is 2.5. Sample includes all eligible schools.

Figure A.2 Proportion treated by forcing variable



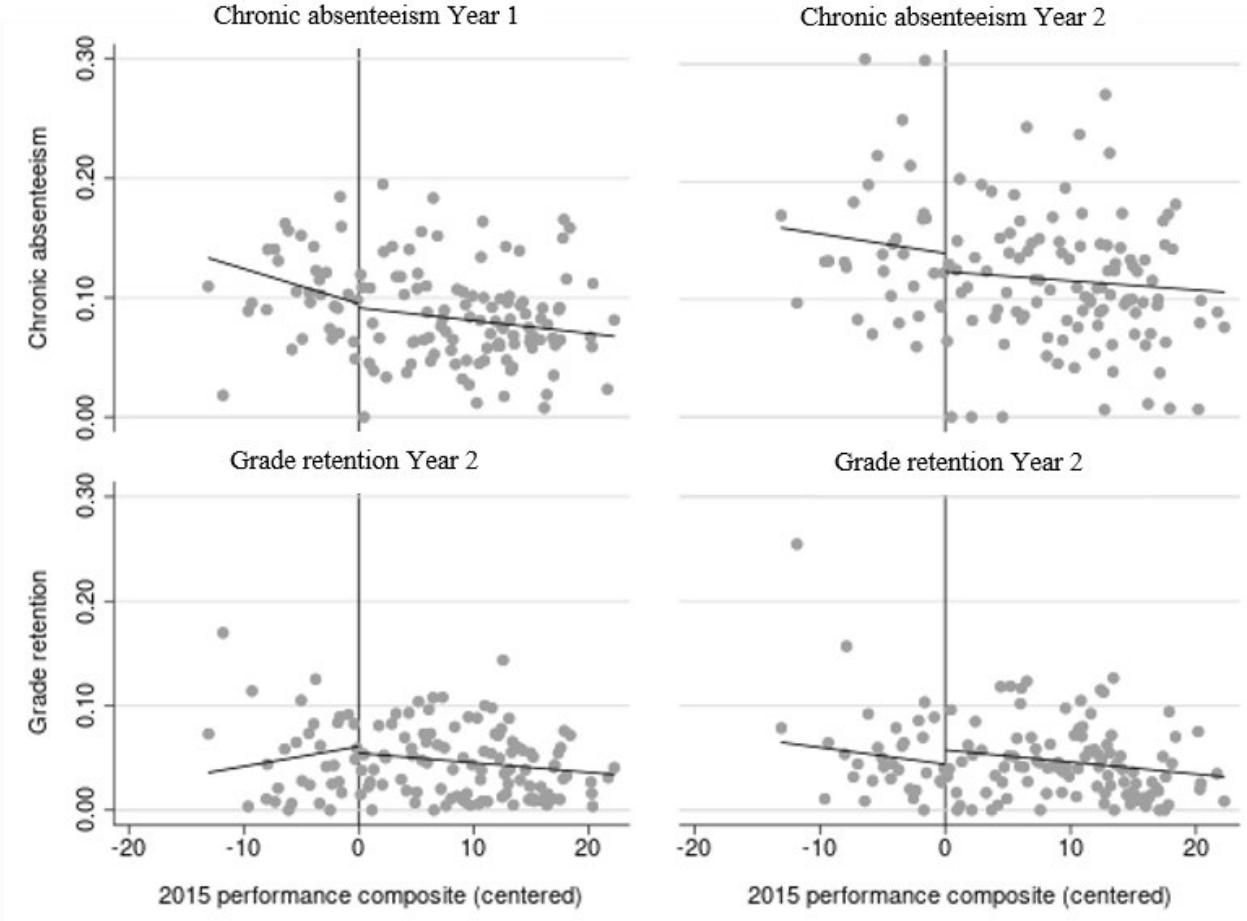
NOTE: Markers represent bin averages. Bin width is 2.5. Marker sizes weighted by number of schools in bin.

Figure A.3. The effects of NCT on early literacy and reading comprehension across the full sample



NOTE: Schools were divided into groups or "bins" based on their 2015 performance composite. Then, for each bin, the average of the school-level mean outcome (early literacy or reading comprehension) was calculated and plotted as a dot on the figure. Line is linear fit.

Figure A.4. The effects of NCT on chronic absenteeism and grade retention across the full sample



NOTE: Schools were divided into groups or "bins" based on their 2015 performance composite. Then, for each bin, the average of the school-level mean outcome (chronic absenteeism or grade retention) was calculated and plotted as a dot on the figure. Line is linear fit.

Appendix B

Table B.1 Estimates on early literacy, reading comprehension, chronic absenteeism, & grade retention by grade

Panel A. Kindergarten

	Year 1			Year 2		
	(1) CCT	(2) 150% CCT	(3) 200% CCT	(4) CCT	(5) 150% CCT	(6) 200% CCT
Early literacy	-0.267**	-0.165*	-0.106	-0.059	0.016	0.096
	(0.0867)	(0.0708)	(0.0615)	(0.0929)	(0.0747)	(0.0635)
N	*9398*	*9398*	*9398*	*9238*	*9238*	*9238*
N within bandwidth	*1303*	*2120*	*3024*	*1335*	*2051*	*2911*
Reading comprehension	-0.116	-0.032	-0.043	0.282**	0.239**	0.148*
	(0.0955)	(0.0778)	(0.0673)	(0.0931)	(0.0778)	(0.0667)
N	*9267*	*9267*	*9267*	*8886*	*8886*	*8886*
N within bandwidth	*1363*	*2179*	*2991*	*1303*	*1982*	*2787*
Chronic absenteeism	0.066*	0.070**	0.042	0.067*	0.064*	0.041
	(0.0304)	(0.0253)	(0.0219)	(0.0329)	(0.0282)	(0.0252)
N	*11127*	*11127*	*11127*	*10849*	*10849*	*10849*
N within bandwidth	*1589*	*2495*	*3612*	*1524*	*2420*	*3465*
Grade retention	0.057*	0.058**	0.038*	-0.022	0.006	0.013
	(0.0226)	(0.0192)	(0.0165)	(0.0281)	(0.0216)	(0.0175)
N	*11127*	*11127*	*11127*	*10849*	*10849*	*10849*
N within bandwidth	*1589*	*2495*	*3612*	*1524*	*2420*	*3465*
Bandwidth	2.9	4.3	5.7	2.9	4.3	5.7
N schools below cutoff	14	22	27	14	22	27
N schools above cutoff	12	19	29	12	19	29

Panel B. Grade 1

	Year 1			Year 2		
	(1) CCT	(2) 150% CCT	(3) 200% CCT	(4) CCT	(5) 150% CCT	(6) 200% CCT
Early literacy	-0.230**	-0.032	0.010	-0.041	0.044	0.051
	(0.0799)	(0.0644)	(0.0543)	(0.0822)	(0.0686)	(0.0577)
N	*9953*	*9953*	*9953*	*9251*	*9251*	*9251*
N within bandwidth	*1436*	*2250*	*3240*	*1241*	*1922*	*2848*
Reading comprehension	-0.381***	-0.010	0.021	0.022	0.212**	0.122*
	(0.0692)	(0.0570)	(0.0487)	(0.0831)	(0.0687)	(0.0566)
N	*9864*	*9864*	*9864*	*8578*	*8578*	*8578*
N within bandwidth	*1500*	*2322*	*3282*	*1151*	*1794*	*2717*
Chronic absenteeism	0.019	0.029	0.028	-0.017	-0.000	-0.009
	(0.0255)	(0.0210)	(0.0183)	(0.0314)	(0.0266)	(0.0230)
N	*11902*	*11902*	*11902*	*11397*	*11397*	*11397*
N within bandwidth	*1760*	*2755*	*3925*	*1597*	*2520*	*3674*
Grade retention	0.070**	0.053**	0.039*	0.058*	0.033	0.021
	(0.0212)	(0.0176)	(0.0154)	(0.0255)	(0.0206)	(0.0174)
N	*11902*	*11902*	*11902*	*11397*	*11397*	*11397*
N within bandwidth	*1760*	*2755*	*3925*	*1597*	*2520*	*3674*
Bandwidth	2.9	4.3	5.7	2.9	4.3	5.7
N schools below cutoff	14	22	27	14	22	27
N schools above cutoff	12	19	29	12	19	29

Panel C. Grade 2

	Year 1			Year 2		
	(1) CCT	(2) 50% CCT	(3) 200% CCT	(4) CCT	(5) 150% CCT	(6) 200% CCT
Early literacy	-0.222***	-0.182***	-0.102**	0.086	0.077	0.098**
	(0.0561)	(0.0455)	(0.0382)	(0.0536)	(0.0433)	(0.0373)
N	*9935*	*9935*	*9935*	*9503*	*9503*	*9503*
N within bandwidth	*1362*	*2150*	*3084*	*1389*	*2175*	*3034*
Reading comprehension	-0.207**	-0.224***	-0.131**	-0.185**	-0.088	-0.020
	(0.0679)	(0.0563)	(0.0475)	(0.0692)	(0.0558)	(0.0475)
N	*10002*	*10002*	*10002*	*8890*	*8890*	*8890*
N within bandwidth	*1522*	*2289*	*3190*	*1368*	*2098*	*2936*
Chronic absenteeism	0.004	0.006	0.010	-0.028	-0.026	-0.004
	(0.0278)	(0.0209)	(0.0170)	(0.0279)	(0.0237)	(0.0209)
N	*11812*	*11812*	*11812*	*11764*	*11764*	*11764*
N within bandwidth	*1750*	*2701*	*3839*	*1690*	*2636*	*3860*
Grade retention	-0.007	-0.011	-0.010	-0.038*	-0.036*	-0.022
	(0.0230)	(0.0179)	(0.0144)	(0.0179)	(0.0140)	(0.0114)
N	*11812*	*11812*	*11812*	*11764*	*11764*	*11764*
N within bandwidth	*1750*	*2701*	*3839*	*1690*	*2636*	*3860*
Bandwidth	2.9	4.3	5.7	2.9	4.3	5.7
N schools below cutoff	14	22	27	14	22	27
N schools above cutoff	12	19	29	12	19	29

Estimates from sharp RD using triangular kernel, linear splines, and heteroskedasticity-robust standard errors. CCT refers to the RD bandwidth selection procedure described by Calonico, Cattaneo, & Titiunik (2014). Early literacy and reading comprehension models are conditioned on beginning-of-year scores, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, and days between beginning and end of year assessments. All models control for school and student covariates. School covariates include minority percentage, economically disadvantaged percentage, per pupil expenditures (PPE) and PPE squared, and enrollment and enrollment squared. Student covariates include gender, race/ethnicity with white as the reference category, disabled, limited English proficient, over-age for grade, and nonstructural transfer in. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$