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

Abstract: Test-based accountability pressures have been shown to result in transferring less effective teachers into untested early grades and more effective teachers to tested grades. In this paper, we evaluate whether a state initiative to turnaround its lowest performing schools reproduced a similar pattern of assigning teachers and unintended, negative effects on the outcomes of younger students in untested grades. Using a sharp regression discontinuity design, we find consistent evidence of increased chronic absenteeism and grade retention in the first year. Also, the findings suggest negative effects on early literacy and reading comprehension in the first year of the reform 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

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

The high-stakes accountability systems that became ubiquitous under the No Child Left Behind Act (NCLB, 2001) and continue under the Every Student Succeeds Act (ESSA, 2015) 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). If student learning and engagement in grades K-2 declines as a result of schools redirecting resources away from untested early grades, the goals of accountability-driven interventions, such as school turnaround in low-performing schools, could be undermined. 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. Specifically, 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 3rd -8th grade accountability system, on student cognitive and non-cognitive 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?

To preview our results that rely upon a sharp regression discontinuity design, we find that the intervention had negative effects on chronic absenteeism and grade retention in the first year of the reform followed by null effects in the second year. Also in the first year of turnaround, we find evidence of negative effects on early literacy and reading comprehension with scores rebounding in the second reform year—though we consider this evidence to be suggestive. In terms of the second research question, schools strategically reassigned low effectiveness teachers to untested courses, but these strategic assignment practices were not significantly more prevalent in treatment than comparison schools. Our findings suggest that school leaders of low-performing schools transferred resources in the form of higher quality teachers in ways that could undermine student learning in early grades.

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 K-2 classrooms—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; Dynarski et al., 2013; 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 3rd through 8th grades. In order to boost school performance and avoid the consequences of being labeled a "failing" school, these accountability pressures incentivize school leaders to disproportionately concentrate resources in tested grades and subjects and away from untested early grades.

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 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, Papay, and Chi (2020) found similar strategic staffing patterns based on principal performance ratings of teachers in one large North Carolina district between 2002 and 2010. These strategic staffing practices are especially prevalent in schools with low accountability grades, 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; 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, Kalogrides, and Loeb (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). In particular, teachers in early grades that do not benefit from the coaching supports that were largely targeted at teachers of tested grades may respond differently to the low-performing label or being assigned to a turnaround intervention because they are subject to the stigma of being classified as low performing without the benefits of instructional supports and additional resources. Research on teacher responses to accountability pressures suggests that teacher job commitment is higher when teachers have an expectation of improvement (Mintrop, 2003). To that end, it is possible that the turnaround designation may undercut morale and teaching effort or quality in early grades by introducing a stigma without counteracting that stigma with supports.

In this study, we seek to extend the literature on the unintended effects of test-based accountability on early childhood investments by examining the effects of a school turnaround intervention on students in untested grades. We do so in the context of the North Carolina Transformation (NCT) initiative, which designated the lowest performing schools—based on standardized test score proficiency—as turnaround schools. These schools received turnaround services from 2015 through 2017. School turnaround represents a specific high-stakes accountability context in which schools that receive low accountability scores are provided with extra support and resources (e.g., needs assessments, school improvement planning, leadership and instructional coaching) from local or state education agencies in order to improve school performance. These turnaround supports can also include intensive actions to disrupt the status quo in low-performing schools, such as replacing the principal and at least half the staff or

turning over the management of the school to a charter management organization. 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 et al., 2016; Zimmer et al., 2017), no studies have analyzed the unintended effects of school turnaround on early-grade student outcomes. We examine these unintended effects on student achievement, non-cognitive outcomes, and teacher assignments. Student achievement 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 student non-cognitive 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, a pattern previously documented in North Carolina during the NCLB era (Fuller & Ladd, 2013; Kraft et al., 2020).

This paper proceeds as follows. In the next section, we describe the NCT intervention and associated theory of change. We then review the methods and empirical strategy of our study before turning to a presentation of the results. We end with a discussion of findings, including relevance and implications for future accountability and school turnaround research.

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 focus of the present study. NCT served the state's low-performing schools during the period between Race to the Top (RttT) 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 (DPI) under the direction of their District and School Transformation (DST) unit. Figure 1 graphically displays the theory of change for the NCT 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 (CNA) in which coaches assigned by DST 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. CNA findings were then "unpacked" or discussed with treatment school staff. These 1.5-day unpacking sessions involved reviewing the CNA 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 CNA, although there was variation in when and whether schools received unpackings. Following the CNA and unpacking, the theory of change called for schools to create their School Improvement Plans (SIP) to outline their priorities and goals. Schools then submitted SIPs through an online platform called NCStar, and state coaches provided feedback through the same platform. The CNA, unpacking, and SIP 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 CNA, unpacking, and SIP, coaches were assigned to NCT schools with the goal of building school capacity. School transformation coaches (STCs) worked with principals and instructional

coaches (ICs) worked with teachers. Under NCT, there were no formal or state-mandated coaching requirements; instead, coaches provided tailored supports to principals and teachers based on their needs. On average over the three semesters of coaching from spring 2016 through spring 2017, schools assigned to treatment received 37 instructional coach visits and 19 school transformation coach visits—though there was large variation in the number and content of coaching visits by school, with IC visits ranging from 0 to 79 and STC visits ranging from 0 to 49. Because coaching visits were aligned with the SIP, they were likely to be concentrated in the tested grades and subjects because the performance measure was performance on 3rd through 8th grade high-stakes tests in reading, mathematics and science. Thus, the intervention partly focused on whole school improvement and partly 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.

A study examining the effect of NCT on student test score growth on end-of-grade (EOGs) and end-of-course (EOCs) exams in grades 4 and above found no effect in the first year of the intervention followed by a .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 CNAs 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.

Methods

Data

This study relies on two sources of data. First, 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 (EPIC) containing data on all students, teachers, and schools in North Carolina. We use data from 2014-15 through 2017-18. We merge the administrative data with mCLASS K-2 student literacy data. The mCLASS includes DIBELS, which involves short, one-minute assessments of student phonemic awareness, alphabetic knowledge, and reading fluency, and TRC, which 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. We use mCLASS literacy data from the 2015-16 and 2016-17 school years.

Analytic Sample

The 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 (SPG) of C or above for the 2014-2015 school year, exceeded growth as measured by the state's Educator Value-Added Assessment System (EVAAS), were part of one of the 10 largest school districts in the state or in

Halifax County (which participated in a district-level turnaround during the same time as the NCT intervention), or were designated as a special or charter school. The state of North Carolina 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 end-of-grade 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 the 38 schools scoring below 31.1 being targeted for services.[1] 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.[2] In total, 35 schools that enrolled K-2 students received treatment under the NCT intervention. Sample school characteristics are displayed in Table 1. There were 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

[1] 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.

[2] Eligibility for NCT was a strong predictor of participation in NCT. 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. See Figure A.2 in the appendix for a graphical depiction of the proportion of schools treated by the forcing variable.

grades K-2 and tested subjects in grades 3-8 who taught in a treatment or comparison 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.

Outcome Measures

We estimate the effects of NCT on four student outcomes: early literacy, reading comprehension, chronic absenteeism, and grade retention. 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. While the mClass is intended as a formative assessment, validation research has shown that DIBELS has high predictive validity with a third-grade end-of-grade reading exam (Smith et al., 2020). We standardize the DIBELS and TRC scores by grade, year, and period (i.e., beginning- or end-of-year exam). Chronic absenteeism is a binary indicator that takes a value of 1 when a student is absent for 10 percent or more of enrolled school days. 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.

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. We merge teacher demographic and evaluation data with student course-level data to answer this question. For teacher assignments, we use both assessment and roster data and code a teacher as teaching in a tested grade and subject if she teaches a grade-subject

combination with an end-of-grade (EOG) exam. In this sample, teachers of tested courses include those teaching math or reading to students who take math or reading EOGs in third through eighth grade, or who teach science to students who take science EOGs 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. We measure teacher effectiveness using subject-level value-added scores (Education Value-Added Assessment System, or EVAAS). The state calculates EVAAS scores using EOGs for teachers in tested courses and using mCLASS TRC reading comprehension assessments for K-2 teachers, whose students do not take EOGs. EVAAS scores are available for about 90 percent of teachers in the sample.[3] Teachers receive one of three ratings based on their EVAAS score for a given subject—meet expected growth, exceed expected growth, or do not meet expected growth. These categories are relevant because school leaders receive EVAAS scores and growth categories for individual teachers and can use them to make staffing decisions. We therefore follow these growth categories set by the state and received by school leaders, coding a teacher as "low effectiveness" if she has a EVAAS score of less than -2, "high effectiveness" if she has a EVAAS score greater than 2, and "mid effectiveness" if all EVAAS scores fall within 2 points of the mean. About 20 percent of teachers with EVAAS scores are low effectiveness, 66 percent are mid effectiveness, and 14 percent are high effectiveness.

[3] In the tested sample, 85 percent of teachers have EVAAS scores. In the untested sample, 95 percent have EVAAS scores.

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.[4] 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.

Empirical Strategy

[4] 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.

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 near 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. The model takes the form

$$y_{is} = \beta_0 + \beta_1 I(GLP < 0)_s + \beta_2 f(GLP)_s + \beta_3 I(GLP < 0)_s \times f(GLP)_s + \gamma \boldsymbol{S}'_s + \sigma \boldsymbol{K}'_i + \varepsilon_{is} ,$$

where *y* represents the student outcome (early literacy, reading comprehension, chronic absenteeism, or grade retention) for student *i* in school *s* at the end of the year. *GLP* represents the forcing variable, *I(GLP)* is an indicator for treatment eligibility that takes a value of 1 in schools below the assignment threshold, *f(GLP)* is a flexible function of the distance from the cutoff, the interaction between the treatment eligibility variable and forcing variable allows for a different slope on either side of the cutoff, and ε is an idiosyncratic error term. 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 overreject the null (Cameron & Miller, 2015). $\boldsymbol{S}'$ is a vector of school-level covariates and $\boldsymbol{K}'$ is a vector of student-level covariates including the student's

score on the beginning-of-year exams for early literacy and reading comprehension, respectively. β_1 is the coefficient of interest, representing the estimated discontinuity at the cutoff. To model the effect of NCT around the cutoff, we estimate locally weighted linear regressions using a triangular kernel within the bandwidth calculated using the mean square error (MSE)-optimal bandwidth selection procedure described by Calonico, Cattaneo, & Titiunik (2014), which accounts for the clustered assignment of schools to treatment.

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 comparison 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 takeup. We consider the ITT estimates to be the policy-relevant estimator and they would not be subject to bias arising from differences between schools that complied or did not comply with their original treatment assignment.

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 WWC standards, we run models within a series of alternative bandwidths, including 150% and 200% of the CCT bandwidth.[5] 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 courses based on teacher effectiveness and experience. Specifically, we are interested in whether low effectiveness 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 comparison schools during the study period (t = 2015, 2016, and 2017).[6] To classify teachers as effective or ineffective, we need teachers to teach either an EOG 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 EOGs. The second sample comprises teachers who teach early grade reading (i.e., reading in

[5] 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.

[6] We do not include teachers of untested non-academic subjects (e.g., physical education, music) in in our analysis.

K, 1, or 2). These teachers receive EVAAS scores based on mClass reading comprehension exams.

Using logistic regression, we predict whether tested and untested teachers return to the same school and teach in an untested early grade in grade *t*+1. Teachers who return to the same school and teach in an untested early grade in *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 *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 predict these outcomes separately 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.[7] We run two sets of these models, with the first predicting teacher assignment using teacher effectiveness based on EVAAS scores, and the second predicting 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. The teacher effectiveness model takes the form

$$ln\left(\frac{\Pr(returning\ and\ teaching\ in\ untested\ early\ grade)}{\Pr(leaving\ the\ school\ or\ returning\ and\ teaching\ outside\ K-2\ academic\ subjects)}\right)$$
$$= \beta_0 + \beta_1 NCT_s + \beta_2 Low_i + \beta_3 High_i + \beta_4 NCT_s \times Low_i + \beta_5 NCT_s \times Mid_i$$
$$+ \beta_6 NCT_s \times High_i + \gamma \boldsymbol{S}'_s + \sigma \boldsymbol{T}'_i + \varepsilon_{is}$$

[7] Teachers are more likely to be classified as effective using TRC scores than EOG scores, so teachers who are already in untested grades are disproportionately classified as highly effective relative to teachers in tested grades and subjects.

predicting the log odds of returning to the same school in an untested early grade (relative to either leaving the school or returning and teaching outside K-2 academic subjects) as a function of treatment assignment (*NCT*), teacher effectiveness category (*Low* and *High*, with mid-effectiveness as the omitted category), interactions between treatment and effectiveness category (*NCT*x*Low*, *NCT*x*Mid*, and *NCT*x*High*), vectors of school (***S'***) and teacher-level (***T'***) 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 would be apparent in β_2 and β_3, while evidence of differential strategic staffing practices in treatment schools would be apparent in β_4 and β_6. Positive estimates on β_2 and β_4 would provide evidence of strategic staffing with respect to low effectiveness teachers, while negative estimates on β_3 and β_6 would provide evidence of strategic staffing with respect to high effectiveness teachers. In particular, a positive estimate on β_2 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 β_4 would suggest that strategic assignment of low effectiveness teachers to untested grades was more prevalent in NCT schools than comparison schools. A negative estimate on β_3 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 β_6 would suggest that strategic assignment of highly effective teachers to untested grades was more prevalent in NCT than comparison schools. If schools were not engaging in strategic staffing to

the potential detriment of untested early grades, we would expect to see insignificant 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.

Results

The results section proceeds as follows. We first discuss the effects of the NCT intervention on cognitive outcomes, followed by the intervention effects on noncognitive outcomes. We then describe our findings on the strategic reassignment of teachers in untested early grades.

Cognitive outcomes

We find evidence in our main models 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 2016 and the second for 2017. In each graph, the horizontal axis represents the 2014-15 school performance composite centered at the eligibility threshold. 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 2016 are apparent in the discontinuity between the lines to the left and right of the cutoff.

Figure 2 ABOUT HERE

The discontinuities at the cutoff are less consistent across outcomes and less pronounced in 2017. We therefore turn to our regression results to interpret both sets of estimates in Table 2. Column 1 shows the estimates within the preferred bandwidth for 2016. As the graphical results depict, the RD finds 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 .2 standard deviations lower in NCT schools than in control schools. 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 and 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.

Columns 4-6 show that treatment schools rebounded somewhat in the second year of services. Columns 4 shows marginally significant positive effects in 2017 of about .08 to .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 2017 estimates were robust to the local randomization RD within most balanced windows (Appendix A, Table A.2).

Reading comprehension and early literacy effects are qualitatively similar across grade levels, with the most consistent Year 1 negative effects in second grade and the strongest and most consistent Year 2 positive effects in reading comprehension in kindergarten. We provide these results in Appendix B.

Table 2 ABOUT HERE

Noncognitive outcomes

We turn next to the effect of NCT on chronic absenteeism and grade retention, shown in Figure 3. The discontinuity between the two linear splines in 2016 shows that NCT schools—i.e., those schools to the left of the cutoff—had higher values of both chronic absenteeism and grade retention in the first year of the intervention. Table 2 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 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.

The effects on grade retention were largely concentrated in kindergarten and first grade, as we show 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. There was no effect on grade retention in second grade. The effects on chronic absenteeism were strongest and most consistent in kindergarten.

Figure 3 ABOUT HERE

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 comparison 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 3, 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 2016—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 comparison 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 EOGs, in treatment schools who moved to untested early grades. This finding shows that treatment schools were retaining all their highly effective teachers in tested 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 2017, NCT schools had no highly effective teachers in tested grades who were reassigned to untested early grades.

Table 3 ABOUT HERE

Table 3, Panel B shows that across the full sample, novice teachers in untested early grades were less likely than experienced teachers to remain in untested courses the following year (see Row 3, Columns 1 and 3)—suggesting that on average, treatment and comparison schools were not hiding their inexperienced teachers in untested courses. 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 2017, when early grade student achievement rebounded somewhat in NCT schools.

Discussion

In this paper, we find evidence that a school turnaround initiative largely focused on improving instruction in tested grades had 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—are 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).

Meanwhile, the negative effects in early literacy and reading comprehension rebounded in the second year, though not to the extent that they dipped in Year 1.

One potential mechanism that may help to explain the 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 concentrated in 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. Additionally, the Year 2 reading comprehension increase was largest among kindergarteners—students who would not have been exposed to the first year of treatment.

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. 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 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 imhibit 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 soley 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 comparison 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 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

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

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

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.

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

Cohen-Vogel, L. (2011). "Staffing to the Test": Are Today's School Personnel Practices Evidence Based? *Educational Evaluation and Policy Analysis Fall XXXX*, *XX*(X), 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

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

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

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

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

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

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., & 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. In *EdWorkingPapers.com*. 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

Marks, J., & Holly, C. (2019). *Guiding Principles for Improving North Carolina's Lowest-Performing Schools*. https://stateboard.ncpublicschools.gov/resources/other-reports/36002nctguidingprinciplesbriefbrochure.pdf

Mintrop, H. (2003). The Limits of Sanctions in Low-Performing Schools. *Education Policy Analysis Archives*, *11*(0), 3. https://doi.org/10.14507/epaa.v11n3.2003

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/

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *The American Economic Review*, *94*(2), 247–252.

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

Standards Handbook (Version 4.0). (2017).

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

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

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. School sample characteristics conditional on forcing variable

	Treatment	Comparison	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	9539.01	9694.70	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 EVAAS	-3.55	-3.25	0.935
N	38	137	

Estimates from sharp RD with covariate listed in row as outcome and triangular kernel.

Table 2. ITT estimates on early literacy, reading comprehension, chronic absenteeism, & grade retention

	2016			2017		
	(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 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 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 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 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. 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 3. Logistic regression estimates of strategic staffing by teacher effectiveness score and experience across treatment and comparison schools

Panel A. Teacher effectiveness score

Odds ratio of teaching in untested grade in *t*+1	2016		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

[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

Panel B. Teacher experience

Odds ratio of teaching in untested course in *t*+1	2016		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

Figures

Figure 1. North Carolina Transformation Theory of Change

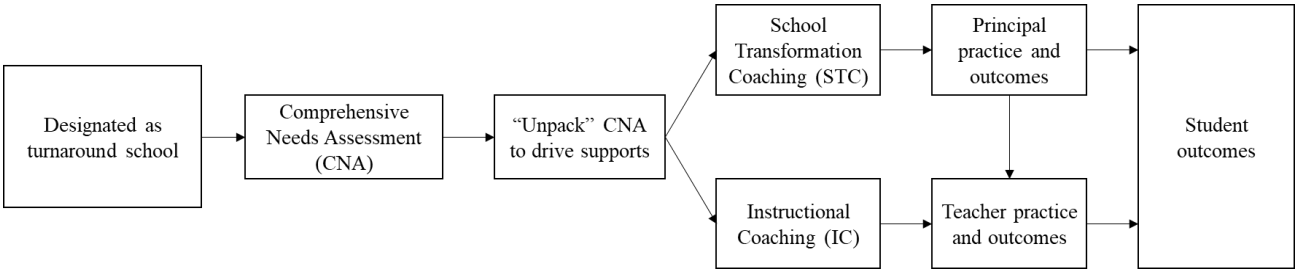
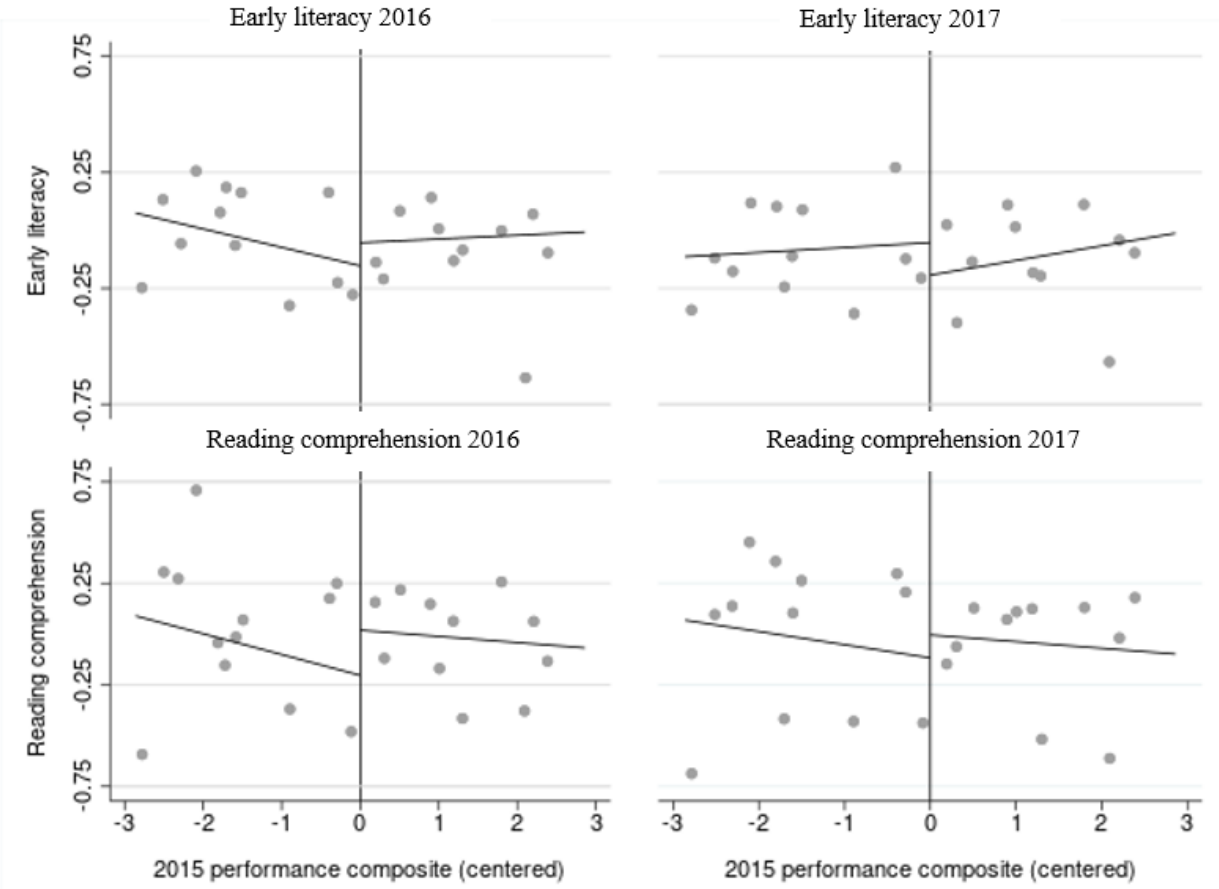
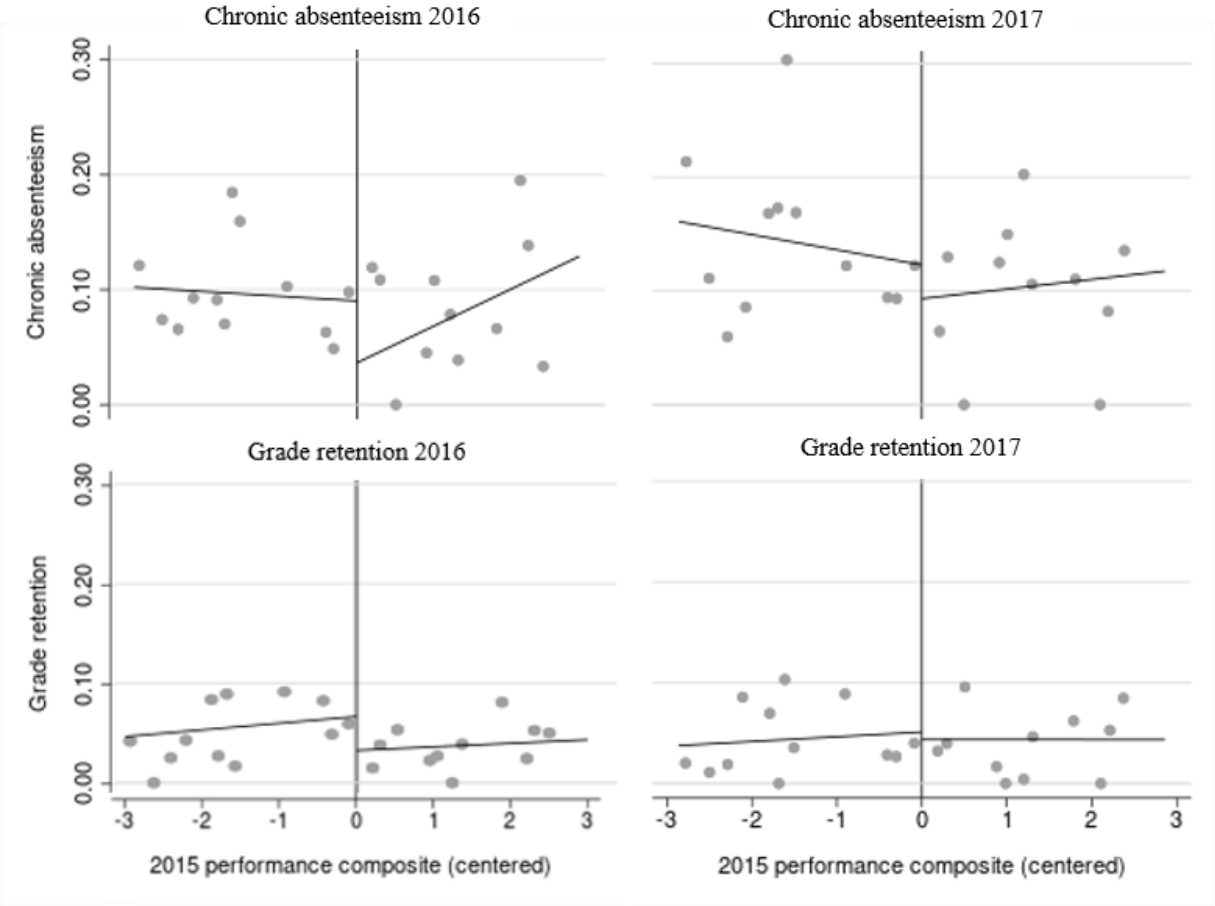


Figure 2. The effects of NCT on early literacy and reading comprehension



NOTE: Markers represent bin averages of school-level means and line is linear fit.

Figure 3. The effects of NCT on chronic absenteeism and grade retention



NOTE: Markers represent bin averages of school-level means and line is linear fit.

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 comparison 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 comparison 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 the 2016 and 2017 years of this study, 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 cognitive outcomes in 2016. 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 cognitive outcomes rebounded in the second year of reform. We also find that chronic absenteeism increased in 2017 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

	2016 & 2017
β_{treat}	0.000
$\beta_{compare}$	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. 2016

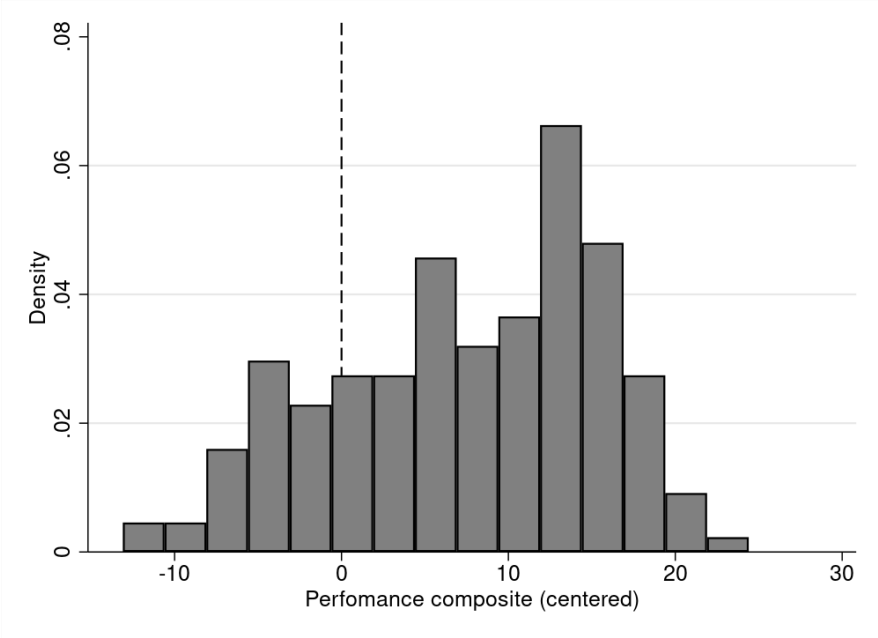
Window length	2.3	2.4	2.5	2.8	2.9	3.2
Early literacy	0.003	0.005	0.017	-0.003	0.041	0.041
Reading comprehension	-0.014	-0.010	0.016	-0.035	0.021	0.021
Chronic absenteeism	0.018*	0.019*	0.018*	0.020*	0.011	0.011
Grade retention	0.020**	0.019**	0.015*	0.014*	0.008	0.008

Panel B. 2017

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

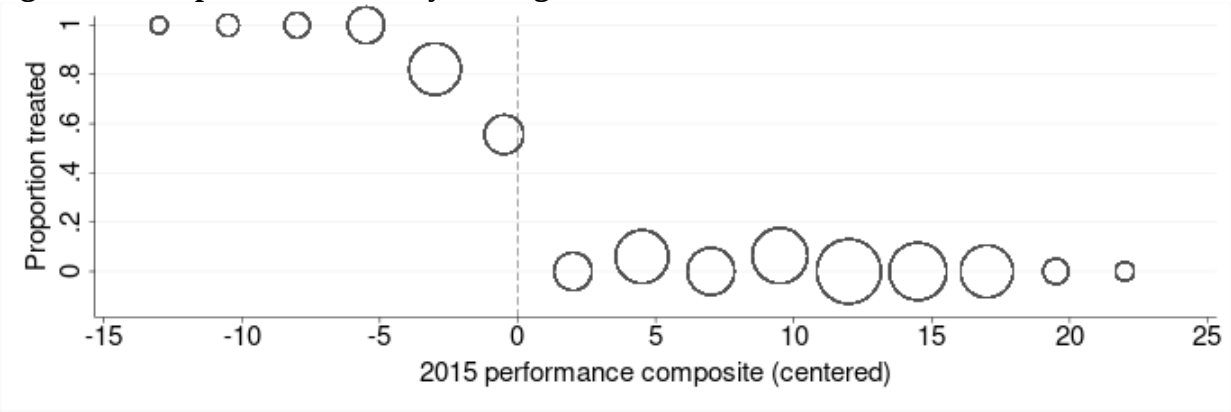
NOTE: Window length represents length on either side of the cutoff. For example, 2.3 runs from -2.3 to +2.3. Estimates from local randomization RD using uniform kernel. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A.1 Graphical integrity of the forcing variable



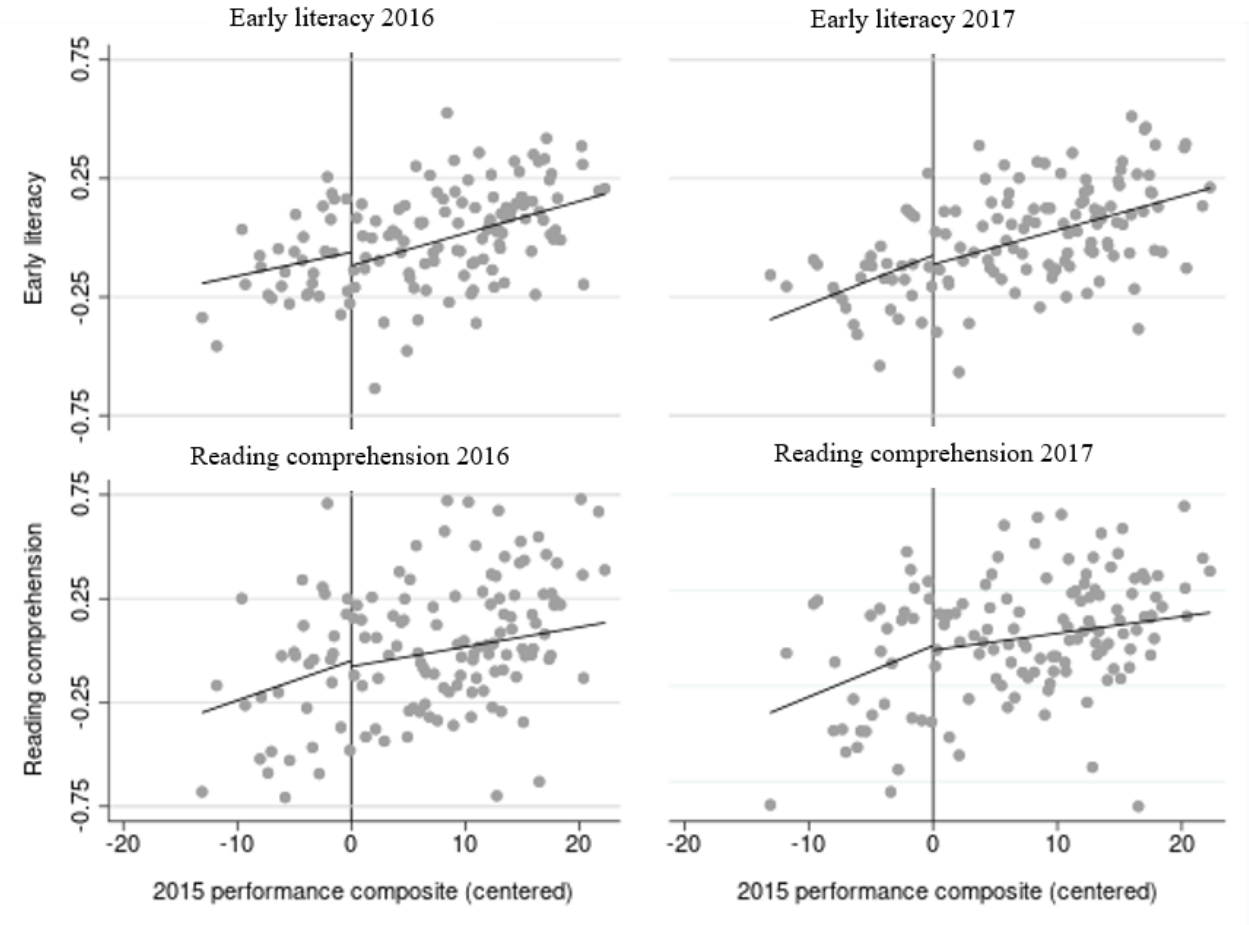
NOTE: Bin width is 2.5. Includes all eligible schools

Figure A.2 Proportion treated by forcing variable



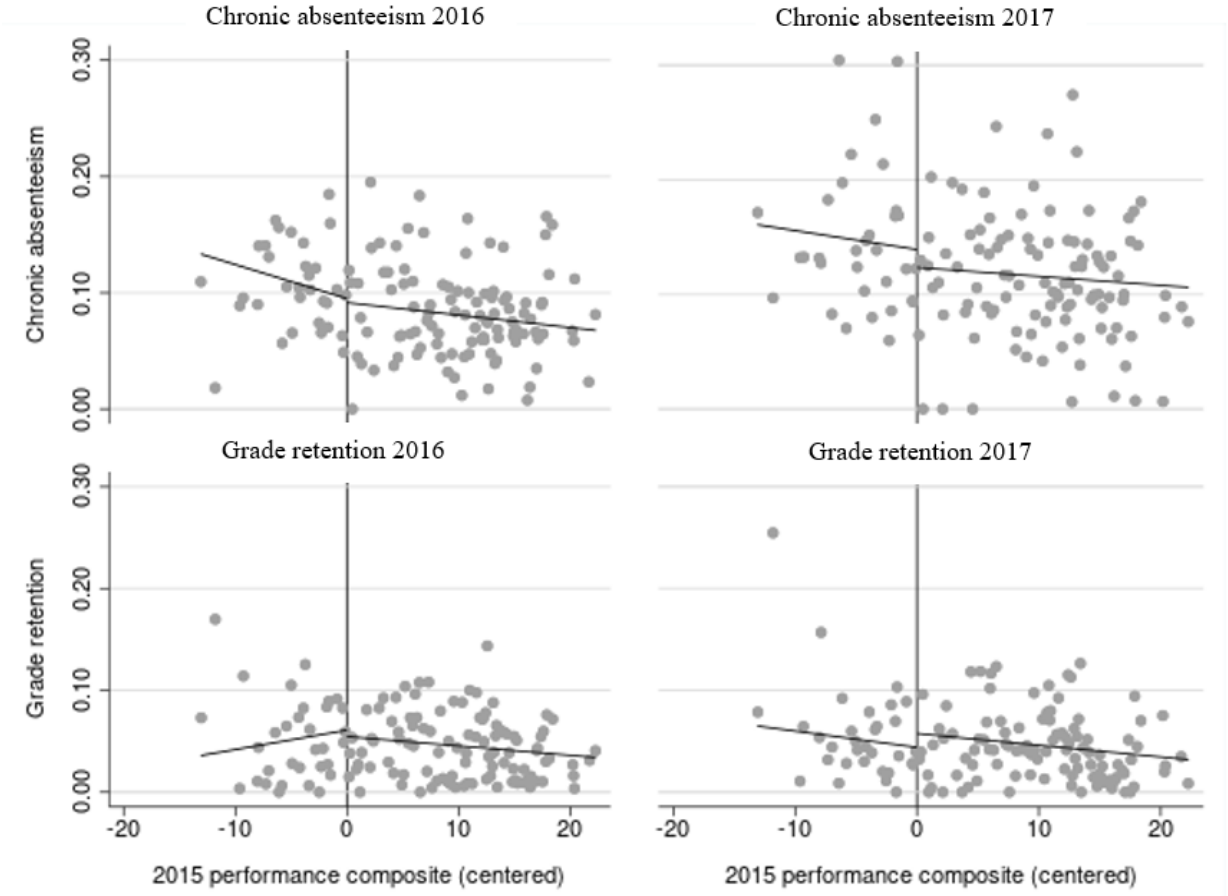
NOTE: Markers represent bin averages. Bin width is 2.5. Marker sizes weighted by number of schools in bin.

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



NOTE: Markers represent bin averages of school-level means and line is linear fit.

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



NOTE: Markers represent bin averages of school-level means and line is linear fit.

Appendix B

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

Panel A. Kindergarten

	2016			2017		
	(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

	2016			2017		
	(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

	2016			2017		
	(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. 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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$