ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-103

The Next Generation of State Reforms to Improve their Lowest Performing Schools: An Evaluation of North Carolina's School Transformation Initiative

Gary T. Henry
Vanderbilt University

Erica Harbatkin
Vanderbilt University

In contrast to prior federally mandated school reforms, the Every Student Succeeds Act (ESSA) allows states more discretion in reforming their lowest performing schools, removes requirements to disrupt the status quo, and does not allocate substantial additional funds. Using a regression discontinuity design, we evaluate a state turnaround initiative aligned with ESSA requirements. We find the effect on student test score growth was not significant in year one and -0.13 in year two. Also in year two, we find that teachers in turnaround schools were 22.5 percentage points more likely to turn over. Teacher turnover appears to have been voluntary rather than the result of strategic staffing decisions.

VERSION: July 2019

Suggested citation: Henry, Gary T., and Erica Harbatkin. (2019). The Next Generation of State Reforms to Improve their Lowest Performing Schools: An Evaluation of North Carolina's School Transformation Initiative. (EdWorkingPaper: 19-103). Retrieved from Annenberg Institute at Brown University: http://www.edworkingpapers.com/ai19-103

The Next Generation of State Reforms to Improve their Lowest Performing Schools: An Evaluation of North Carolina's School Transformation Initiative

Gary T. Henry

Erica Harbatkin

Vanderbilt University

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E150017 to Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

In contrast to prior federally mandated school reforms, the Every Student Succeeds Act (ESSA) allows states more discretion in reforming their lowest performing schools, removes requirements to disrupt the status quo, and does not allocate substantial additional funds. Using a regression discontinuity design, we evaluate a state turnaround initiative aligned with ESSA requirements. We find the effect on student test score growth was not significant in year one and -0.13 in year two. Also in year two, we find that teachers in turnaround schools were 22.5 percentage points more likely to turn over. Teacher turnover appears to have been voluntary rather than the result of strategic staffing decisions.

The Next Generation of State Reforms to Improve their Lowest Performing Schools: An Evaluation of North Carolina's School Transformation Initiative

The mandate for continuous support and improvement of each state's lowest performing schools along with the accountability requirements in the Every Student Succeeds Act (ESSA, 2015) will ensure that every state will continue to identify and attempt to reform its lowest performing schools into the foreseeable future. State turnaround interventions under prior federal programs, School Improvement Grants (SIG) and Race to the Top (RttT), have shown some evidence of positive effects on student outcomes (Carlson & Lavertu, 2018; Dee, 2012; Papay & Hannon, 2018; Sun, Penner, & Loeb, 2017; Zimmer, Henry, & Kho, 2017), although some studies have found negative or null effects (Dickey-Griffith, 2013; Dragoset et al., 2017; Heissel & Ladd, 2018; Henry, Guthrie, & Townsend, 2015). In many of the turnaround efforts that have been shown to be effective, strategic staffing—which involves recruiting, hiring, developing and retaining high quality teachers—seems to have played a role in successful turnaround, as we discuss later.

Central to all four federal turnaround models permitted under RttT and SIG were actions to disrupt the status quo, such as replacing the principal and at least half the staff or turning over the management of the school to a charter management organization. Under ESSA, the four federally mandated turnaround models have faded into the past along with additional dedicated funding for turnaround (provided during RttT and SIG through the American Recovery and Reinvestment Act, or ARRA). The school reform interventions implemented under No Child Left Behind (NCLB) waivers, which did not necessarily follow federally prescribed models and were not supported by ARRA funds, yielded less consistent effects on student achievement than RttT- and SIG-funded interventions, with just one producing positive effects and three null

effects (Bonilla & Dee, 2017; Dee & Dizon-Ross, 2019; Dougherty & Weiner, 2017; Hemelt & Jacob, 2017, 2018).

Turnaround under ESSA will share more in common NCLB waivers and similar state-initiated reforms than RttT and SIG for two reasons. First, states will have flexibility in how they improve their lowest performing schools rather than being required to follow a federally prescribed model. Second, states will undertake turnaround without the infusion of additional federal funds that characterized RttT and SIG reforms. One state-initiated reform operating in this context was the North Carolina Transformation (NCT) initiative, which began in 2015 after the state's services under RttT ended. This study examines the effects of this new round of school support on student achievement and teacher turnover. We ask three research questions:

1. What is the effect of the efforts to improve the lowest performing schools on student achievement?
2. What is the effect of the efforts to improve the lowest performing schools on teacher turnover?
3. Did the reform schools hire more effective replacement teachers than they lost?

By way of preview, relying upon a rigorous regression discontinuity design, we find negative effects on student achievement gains and increased teacher turnover in the second year of services. These findings may serve as a cautionary tale for how states support their lowest performing schools under ESSA.

School Turnaround

Prior research has shown substantial heterogeneity in the effects of whole school reform efforts. Empirical evaluations of the first federally funded whole school improvement program, the Comprehensive School Reform Demonstration program (CSRD), found CSRD programs did

not produce positive average effects on student achievement and in fact may have led to lower math performance among black and Hispanic students (Gross, Booker, & Goldhaber, 2009). In 2008, SIG and RttT introduced school turnaround to the school improvement toolkit, with the distinction that turnaround would create dramatic and rapid change in chronically low-performing schools (Herman et al., 2008; Peurach & Neumerski, 2015). Unlike the prior incremental school reforms under CSRD, RttT and SIG required specific practices for disrupting the status quo as part of federally mandated turnaround models. These intentional disruptions included practices such as replacing the principal, replacing at least 50 percent of staff, or restarting the school under new management to allow complete staff replacement (see, e.g., Zimmer et al., 2017). Turnaround efforts funded through RttT and SIG as well as reforms following similar models in Massachusetts, Tennessee (local Innovation Zones), Ohio, and California—many of which included substantial staff replacement and often practices aimed at recruiting and retaining effective teachers—produced strong positive effects on student achievement (Carlson & Lavertu, 2018; Dee, 2012; Papay & Hannon, 2018; Schueler, Goodman, & Deming, 2016; Strunk, Marsh, Hashim, Bush-Mecenas, & Weinstein, 2016; Sun et al., 2017; Zimmer et al., 2017). The mostly positive effects have largely dominated the conversation about turnaround under RttT and SIG, but the average and local average treatment effects mask substantial heterogeneity within interventions. Turnaround in North Carolina and Texas produced mixed effects (Dickey-Griffith, 2013; Heissel & Ladd, 2018; Zimmer et al., 2017), and some of the interventions yielding positive effects also produced null or negative effects in particular contexts (Carlson & Lavertu, 2018; Strunk et al., 2016; Zimmer et al., 2017, Achievement School District).

Heterogeneity of effects of school reform models continued under NCLB waivers, with fewer positive effects than RttT and SIG. One study found positive effects on student achievement, which the authors attributed to Kentucky's focus on reducing achievement gaps combined with a clearly articulated set of reform activities from the state (Bonilla & Dee, 2017), while reforms in Michigan, Rhode Island, and Louisiana produced either null or negative effects on student achievement (Dee & Dizon-Ross, 2019; Dougherty & Weiner, 2017; Hemelt & Jacob, 2017, 2018).

The mixed effects of interventions under both RttT/SIG and NCLB waivers underscore three important conclusions about school reform. First, recruiting and retaining effective teachers appears to be a key strategy for achieving and sustaining turnaround. Second, successfully shifting the climate and daily operations of an underperforming school may require some disruption of the status quo. And finally, the impacts of school reform interventions are not universally positive or even neutral—these interventions have the potential to do harm, as they did in some schools in Los Angeles, Rhode Island, North Carolina, Texas, and Michigan.

This paper proceeds as follows. In the next section, we describe the intervention and theory of change under NCT and provide some context on implementation. We then describe the sample, data, and empirical strategy, followed by the findings, a series of validity checks, and tests of alternative explanations for the pattern of effects. We conclude with a discussion of the relevance and limitations of these findings for future school turnaround.

North Carolina Transformation Initiative

NCT began during the 2015-16 academic year and was implemented in 75 low-performing schools over two academic years. NCT schools received coaching and support services directly from the state Department of Public Instruction (DPI), which had carried out

two prior rounds of school turnaround interventions. The first, a court-ordered turnaround that prescribed a set of interventions—for example, a freshman academy for high schools—along with instructional and school leadership coaching, included 128 low-performing schools from 2006 through 2010 (Thompson, Brown, Townsend, Henry, & Fortner, 2011). The second was Turning Around the Lowest Achieving Schools (TALAS), the state's RttT turnaround intervention, which focused on reforming 118 schools under the closure (12 schools), transformation (93 schools) and turnaround (14 schools) models through direct service provision from the state Department of Public Instruction (DPI) (Henry et al., 2015). Under TALAS, schools received district-level, school-level, and instructional coaching from about 150 coaches (Henry, Campbell, Thompson, & Townsend, 2014). All schools in the bottom 5 percent of the state based on the 2009-10 proficiency rate received services. When services ended, a leaner DPI set out to continue its work in a smaller group of low-performing schools. An early adopter of turnaround because of the 2006 court order, North Carolina continued its turnaround efforts without the federal pressures that motivated waiver-based reforms during the same time period.

NCT followed a similar direct services model to TALAS but the selection process excluded schools in the 10 largest districts in the state. As a result, NCT schools were largely rural and, on average, higher performing than TALAS schools. NCT also didn't include require implementation of one of the four federal turnaround models or the federally recommended practices. The NCT theory of action as depicted in Figure 1 began with a Comprehensive Needs Assessment (CNA) in which DPI staff would spend two days at treatment schools collecting data through classroom observations, interviews, and focus groups. State staff then produced a report and shared it with the principal, who in turn could choose to share it with school staff. The state prioritized conducting CNAs in NCT schools that had not received one in the three years prior to

the intervention. Of the 75 NCT schools, 84 percent received a CNA immediately prior to this round of school reform or during the two academic years in which services were delivered (See Table 1 for timing of CNAs).

Figure 1 ABOUT HERE

Following the CNA, the NCT model called for an "unpacking" in which state facilitators discussed CNA findings with school staff. There were three key elements of this 1.5-day unpacking process: (1) the facilitators reviewed the full CNA report with attendees, (2) the facilitators and school staff carried out a "root-cause analysis" in which they sought to uncover the underlying causes of the issues identified in the CNA, and (3) the facilitator and school staff engaged in a planning activity that involved visually mapping the school improvement process moving forward. Unpackings generally occurred during the summer following the school year of the CNA, although there was variation in when and whether schools received unpackings.

Table 1 ABOUT HERE

The CNA and unpacking were intended as the springboard from which school turnaround would occur. All low-performing schools in North Carolina were required to submit a School Improvement Plan (SIP) in which priority areas and goals were intended to be based in part on CNA findings. Schools submitted plans through an online tool called NCStar in which school improvement teams were asked to regularly update progress toward school goals. While all low-performing schools submitted these plans and received feedback from the state, NCT schools received feedback on their plans from their coaches. Other low-performing schools received feedback from state staff who may have not been familiar with the schools.

The CNA, unpacking, and SIP comprised the foundation for turnaround. This framework parallels ESSA requirements, which call for districts to work with low-performing schools to

develop a comprehensive support and improvement plan using data from a school-level needs assessment and for the state to monitor school progress on that plan. The core of the intervention was the coaching that followed, with the goal of building school capacity through coaching. School transformation coaches worked directly with principals, while instructional coaches worked with teachers. NCT was intended as a tailored intervention in which coaches were responsive to school, principal, and teacher needs. DPI assigned coaches based on school needs and internal capacity. Not all schools received both school transformation and instructional coaching, and there was wide variation in the number, content, and structure of the visits. District-level coaches also served central office staff in the 43 treatment schools situated in low-performing districts.

Table 2 ABOUT HERE

The intervention did not closely mirror any of the four previous federal school turnaround models, which all contained at least one element intended to disrupt the status quo—either through changes to staff (e.g., replacing at least 50% of teachers), management (e.g., replacing the principal), or governance (e.g., state takeover)—and provided resources to build educational infrastructure. Instead, the NCT theory of change focused on building staff capacity and gave districts autonomy to transform their low-performing schools using locally developed strategies. In its focus on instructional quality, NCT, like RttT and SIG, recognized the importance of highly effective teachers to school turnaround, but did not include a focus on recruitment and retention of effective teachers as the turnaround and transformation models had done. Rather, it focused resources on developing existing staff—similar to the earlier CSRD models. While NCT served the state's low-performing schools during the period between RttT and ESSA, the model aligns more closely with ESSA's flexible approach to school turnaround than with the

prescriptive turnaround models. This evaluation can therefore help to inform state turnaround policy under ESSA, under which states have the flexibility to implement school turnaround interventions that look like NCT.[1]

Sample

The sample includes all North Carolina schools that the state Board of Education and state superintendent of schools determined were eligible for treatment based on data from the 2014-15 school year. Schools were excluded from eligibility for services if they had a school performance grade of C or above for the 2014-15 school year, exceeded growth, were situated in one of the 10 largest school districts in the state, or in Halifax County, which was targeted for a district-level turnaround from 2009-10 through 2016-17. Special schools, charter schools, and freshman academies were also excluded. In total, 331 schools were eligible for services, and 78 were targeted for treatment. Noncompliance occurred on both sides of the treatment cutoff because state officials did not serve schools without district agreement. In some cases, district officials requested that the state deliver services to a school above the cutoff rather than the school selected or requested that a particular school be served in addition to the targeted schools. Sixty-nine of the 78 schools below the cutoff complied with their assignment, nine below the cutoff declined, and six schools above the cutoff received services.

Of the 78 schools below the assignment threshold, 72 were rural, five were in towns, and one was in a city. On average, treatment schools had higher rates of minority and low-income students, higher rates of novice teachers, higher per pupil spending, and lower enrollment than other eligible schools, which were higher performing, as Table 3 shows. The state identified schools proportionally by level based on the eligible population of schools, with 38 elementary, 28 middle, and 12 high schools assigned to treatment.

Table 3 ABOUT HERE

Data

This analysis draws from a longitudinal database of statewide administrative data maintained by the University of North Carolina-Chapel Hill's Educational Policy Initiative at Carolina (EPIC). The database contains data on all students, teachers, principals, and schools in North Carolina public schools. Our analysis uses student-level data to estimate the effect of NCT on student achievement and teacher-level data to estimate the effect on teacher turnover.

Outcome measures

We estimate the effect of NCT on end-of-grade (EOG) and end-of-course (EOC) test score growth. Students in North Carolina take math and reading EOGs each year in third through eighth grade, science EOGs in fifth and eighth grade, and EOCs in Math 1, English II, and Biology. Exams are administered in the final 10 instructional days of the school year for year-long courses and the final five instructional days of fall semester for half-year block EOC courses taken in the fall. We operationalize teacher turnover as leaving the school, either to move to another school or leave North Carolina public schools altogether. Teacher turnover is measured during and at the end of the school year, so a teacher who does not return to her original school in the 2016-17 school year would be counted as having turned over in 2015-16.

Assignment variable

The state assigned schools to receive services based on the 2014-15 school performance composite, a measure that represents the EOG and EOC exam passage rate (abbreviated below as GLP, for grade-level proficiency). To account for differences in passage rates by exam and ensure the proportion of treated elementary, middle, and high schools roughly matched the eligible sample's proportion of schools at each level, the state set separate cutoffs for elementary,

middle, and high schools. The cutoff was 31.1 for elementary schools, 33.8 for middle schools, and 26.0 for high schools. Schools below these thresholds were targeted for services. For the analysis, we center the performance composite at the threshold by school level.

Teacher effectiveness

To explore whether teacher turnover represents strategic deselection of the lowest performing teachers, we draw from two lagged measures of teacher effectiveness. Subject-level value-added scores (Education Value-Added System, or EVAAS) provide a measure of teacher effectiveness for teachers of tested grades and subjects, while the teacher's evaluation ratings as measured by the North Carolina Educator Effectiveness System (NCEES) are available for teachers of tested and untested grades and subjects. We use EVAAS scores calculated from EOCs and EOGs, as well as mClass reading assessments in kindergarten through third grade. About one-third of teachers in the sample have lagged scores in each outcome year. Teachers receive one of three ratings based on their EVAAS score for a given subject: they *meet expected growth* if they are within 2 points of predicted growth on the EVAAS scale, *exceed expected growth* at more than 2 points above, and *do not meet expected growth* at more than 2 points below. We use these cutoffs to place teachers in effectiveness categories. Specifically, we code a teacher as "highly effective" if she has a lagged EVAAS score that exceeds expected growth, "low effectiveness" if she has a lagged EVAAS score that does not meet expected growth, and "mid effectiveness" if all EVAAS scores fall in the meets expected growth category.[2]

NCEES includes five standards: (1) teacher leadership, (2) establishing a respectful learning environment for diverse students, (3) content knowledge, (4) facilitate learning for students, and (5) reflecting on practice. Teachers receive ratings of 1 to 5 on each rating, with 1 being the lowest rating a teacher can receive and 5 the highest. Because teachers with more than

three years of experience are only required to be evaluated on standards 1 and 4, we draw the NCEES measures from these two standards. We observe lagged NCEES ratings on each of these standards for about 70 percent of the sample during the outcome years. We generate two different NCEES effectiveness measures—one for standard 1 and one for standard 4. The modal rating in the sample on both measures is a 3. We again place teachers into three effectiveness categories based on these lagged NCEES ratings: "low effectiveness" for teachers with a 1 or 2, "mid effectiveness" for teachers with a 3, and "highly effective" for teachers with a 4 or 5.[3]

Using EVAAS and NCEES, we end up with three categorical measures of teacher effectiveness: high, mid, and low EVAAS; high, mid, and low NCEES standard 1; and high, mid, and low NCEES standard 4. Each has distinct advantages and disadvantages. EVAAS contains the most variation but restricts the sample to just teachers who were in tested grades and subjects the prior year. NCEES captures more of the sample but classifies very few teachers in the low category (about 2% of teachers in the sample).

Covariates

School-level variables include minority percentage, economically disadvantaged percentage, per pupil expenditures (PPE) and PPE squared, average daily membership (ADM) and ADM squared, and school level with elementary as the reference category. Teacher-level variables include female and race with white as the reference category. Student-level variables include female, race with white as the reference category, disabled, academically gifted, limited English proficient, over age for grade, and nonstructural transfer in. We define disabled as a current designation 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 transfer in as a transfer that occurs into the observed school

prior to the maximum grade of the prior school (e.g., transferring into the observed school in 7th grade when the student's prior school went through 8th grade).

Empirical Strategy

Main effects

We estimate the effect of NCT using a regression discontinuity design that exploits the jump in probability of assignment to treatment at the cutoff (Imbens & Lemieux, 2007). We begin with an intent-to-treat (ITT) estimate that takes the form

$$y_{is} = \beta_0 + \beta_1 I(GLP \leq 0)_s + \beta_2 f(GLP)_s + \beta_3 I(GLP \leq 0)_s \times f(GLP)_s + \gamma \boldsymbol{S}'_s + \sigma \boldsymbol{K}'_i + \varepsilon_{is}\,, \quad (1)$$

where *y* is the outcome for student or teacher *i* in school *s*, *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. In a second set of models, we add vectors of school-level covariates, $\boldsymbol{S}'$, and individual-level covariates, $\boldsymbol{K}'$, to increase precision. The individual-level covariates are student level in models predicting student test score growth and teacher level in the teacher turnover models. We also include the student's lagged test score on the right-hand side of the student achievement model. 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.

This ITT analysis is the policy-relevant estimator because it represents the estimated effect of assignment to treatment. However, while eligibility for treatment was a strong predictor of receiving treatment, noncompliance occurred in schools above and below the cutoff. We therefore estimate a treatment on the treated (TOT) estimate using a two-stage least squares (2SLS) model in which we instrument NCT with treatment eligibility. The first stage of the 2SLS model takes the form

$$NCT = \alpha_0 + \alpha_1 I(GLP \leq 0)_s + \alpha_2 f(GLP)_s + \alpha_3 I(GLP \leq 0)_s \times f(GLP)_s + \gamma \boldsymbol{S}'_s + \sigma \boldsymbol{K}'_i + u_{is}, \quad (2)$$

where being in turnaround status (NCT) is a function of a treatment eligibility indicator, *I(GLP≤0)*, that takes a value of 1 if the school was below the treatment threshold; a flexible function of the distance from the cutoff, *f(GLP)*; and an interaction between the two. In the set of models with covariates, we include the vectors of school- and individual-level covariates in the first stage as well. We then estimate the second stage as

$$y_{is} = \beta_0 + \beta_1 (\widehat{NCT})_s + \beta_2 f(GLP)_s + \beta_3 \left(\widehat{NCT_s}\right) \times f(GLP)_s + \pi \boldsymbol{S}'_s + \rho \boldsymbol{K}'_i + \varepsilon_{is}, \quad (3)$$

where the predicted outcome, *y*, for student or teacher *i*, is a function of the predicted *NCT* indicator, and then follows the same format as the first stage. This approach allows us to estimate treatment effects using the schools that complied with their treatment assignment, with β_1 providing an estimated local complier-adjusted treatment effect. The fuzzy RD is our preferred model because it accounts for noncompliance and reflects the estimated treatment effect for compliers.

The TOT estimates would be biased if the instrument failed to meet the exclusion restriction, which requires that the instrument affects the outcome only through the instrumented variable (Angrist, Imbens, & Rubin, 1996). In other words, if districts selected more challenging

schools for treatment and declined treatment for schools that were already improving, the first-stage equation would bias the estimated effect of treatment. The ITT estimates would not be subject to the same bias. Therefore, similar results for the ITT and TOT estimates would suggest the TOT effects are not being driven by bias associated with failure to meet the exclusion restriction.

We stack all subjects in our main student achievement specification but also include separate models for math, reading, and science in the appendix. Because we include the lagged test score on the right-hand side of the equation, the outcome represents one year of subject-level growth for fourth through eighth grade math and reading. For high schools, we measure growth from the eighth-grade EOG exam, which is two years prior for reading and most often one year prior for math. In science, there are two to three years between the lagged score and the outcome score.[4] Because the teacher turnover outcome is a binary indicator for whether the teacher turned over in a given school year, the teacher turnover models are linear probability models in which the RD estimate can be interpreted as the difference in probability of turnover associated with being in a treatment school relative to a control school within the bandwidth.

We also estimate the model within a series of alternative bandwidths, including 50% and 200% of the CCT bandwidth, the optimal bandwidth proposed by Imbens & Kalyanaraman (IK, 2009), 200% of the IK bandwidth,[5] and finally on the full sample of treatment and control schools for which we have implementation data. We cluster standard errors at the school level.[6] Because coaching did not begin until spring 2016—i.e., the second semester of the intervention—we measure the outcomes separately for each year of treatment. The 2016 estimate represents the effect of a single semester of coaching in all schools and a CNA in most schools, while the 2017 estimate represents the effect of a full year of coaching services.[7]

Teacher effectiveness

After estimating the effects of the intervention on student achievement growth and teacher turnover, we conduct an additional analysis to examine the effectiveness of teachers who left the schools and those who entered. Specifically, we are interested in the effects of NCT on teacher turnover and new-to-school teachers by teacher effectiveness category, as defined by the EVAAS, NCEES standard 1, and NCEES standard 4 measures described in the Data section above. To estimate these effects, we run fuzzy RDs predicting two dichotomous outcomes—turnover and being new to school. For each of these outcomes, we run a fuzzy RD with three different treatments—highly effective teachers in NCT schools, mid-effectiveness teachers in NCT schools, and low-effectiveness teachers in NCT schools. In order to estimate within-group differences (e.g., the probability of turnover among low effectiveness teachers in NCT schools relative to low effectiveness teachers in control schools), we also include covariates for high and low teacher effectiveness. Because we have three treatments, we estimate three first-stage models predicting turnaround status within each of the three groups based on teacher effectiveness category. The three first-stage models take the form

$$\begin{aligned} NCT \times TeacherEffectivenessGroup_{is} &= \alpha_0 + \alpha_1 I(GLP \leq 0)_s \times HighlyEffective_i + \alpha_2 I(GLP \leq 0)_s \\ &\times MidEffectiveness_i + \alpha_3 I(GLP \leq 0)_s \times LowEffectiveness_i \\ &+ \alpha_4 f(GLP)_s + \alpha_5 I(GLP \leq 0)_s \times f(GLP)_s + \alpha_6 HighlyEffective_i \\ &+ \alpha_7 LowEffectiveness_i + \varepsilon_{is}\,, \end{aligned} \tag{4}$$

where the first-stage outcome, *NCT x TeacherEffectivenessGroup*, is different for the three first-stage models. Specifically, one first-stage model predicts being a highly effective teacher in a school in turnaround status (*NCTxHighlyEffective*), one predicts being a mid-effectiveness teacher in a school in turnaround status (*NCTxMidEffectiveness*), and one predicts being a low-effectiveness teacher in a school in turnaround status (*NCTxLowEffectiveness*). In other words,

the first stage estimates the joint probability of being in turnaround status and in a particular teacher effectiveness group. These first-stage equations therefore produce three separate estimates to carry into the second stage—one for each of the three interactions representing teachers of high, mid, or low effectiveness in treatment schools. Drawing from each of these three first-stage predictions, we then estimate the second stage as

$$y_{is} = \beta_0 + \beta_1 NCT_s \times \widehat{HighlyEffective}_i + \beta_2 NCT_s \times \widehat{MidEffectiveness}_i + \beta_3 NCT_s \times \widehat{LowEffectiveness}_i + \beta_4 f(GLP)_s + \beta_5 NCT_s \times f(GLP)_s + \beta_6 HighlyEffective_i + \beta_7 LowEffectiveness_i + \varepsilon_{is}\,, \quad (5)$$

where the predicted outcome (turnover or new to school, represented as *y*) for teacher *i* in school *s* is estimated using the same approach as equation 3 except that it provides separate effects by teacher effectiveness category by estimating within-effectiveness-group differences in the probability of turnover or being new to school. Specifically, in the model predicting turnover, β_6 provides the estimated probability of turnover for highly effective teachers and β_1 represents the estimated deviation from that probability for highly effective teachers in treatment schools, while β_7 provides the estimated probability of turnover for low-effectiveness teachers and β_3 represents the estimated deviation from that probability for low-effectiveness teachers in treatment schools. β_2 represents the estimated deviation from B_0 for mid-effectiveness teachers in treatment schools. We estimate these models without additional covariates, though our estimates are robust to inclusion of school- and teacher-level covariates.

Evidence of strategic staffing would be apparent in β_1 and β_3. In the model predicting turnover, a negative and significant estimate on β_1 would provide evidence that treatment schools retained more effective teachers, while a positive and significant estimate on β_3 would provide evidence that less effective teachers left treatment schools, both relative to schools in the control group. In the model predicting new-to-school teachers, a positive and significant estimate on β_1

would provide evidence that treatment schools hired more effective teachers, while a negative and significant estimate on β_3 would provide evidence that treatment schools hired fewer ineffective teachers relative to control schools.

Results

We find consistent evidence that NCT had a negative effect on student achievement growth in 2017. Figure 2 provides a graphical representation of these results within the preferred bandwidth. The vertical distance between the fit lines on either side of the cutoff represents the difference in outcomes associated with being in a school assigned to treatment. The 2017 panel provides graphical evidence of a decrease in student achievement among schools below the cutoff in the second year of services.

Figure 2 ABOUT HERE

Table 4 displays the ITT estimates separately for 2016 (Panel A) and 2017 (Panel B). Model 1, which estimates within the preferred CCT bandwidth, shows that assignment to treatment is associated with a .12 standard deviation decrease in test score growth in the second year of treatment. This result is robust to alternative bandwidths (Models 3–6) and inclusion of covariates (Models 2, 4, and 6).

Table 4 ABOUT HERE

These ITT models provide the policy-relevant estimator, but do not account for noncompliance with treatment assignment, which occurred on both sides of the cutoff. The probability of treatment is high for schools assigned to treatment and low for those not assigned to treatment, but Figure 3 shows that a small proportion of schools below the cutoff did not receive treatment and a small proportion of schools above the cutoff did receive treatment. The

fuzzy RD accounts for this noncompliance by providing the estimated local average treatment effect of NCT for compliers.

Figure 3 ABOUT HERE

The TOT estimates from the fuzzy RD are provided in Table 5. These complier-adjusted treatment effects are similar to the ITT estimates, with an estimated effect of -.13 in 2017 in our preferred model. This similarity provides evidence that the TOT estimates are not biased by a failure to meet the exclusion restriction. Because we are interested in effect estimates that account for noncompliance with treatment assignment and have no evidence that the treatment instrument fails to meet the exclusion restriction, we proceed by showing TOT estimates from the fuzzy RDs moving forward.

Similar to the ITT estimates, the 2017 TOT estimates displayed in Table 5 are consistently negative and significant across bandwidths and with and without covariates. In Appendix Table A-1, we show that the negative effect extends to the full sample of schools in the second year of services. Our results also suggest a negative effect of NCT in 2016 within the narrowest bandwidth. The coefficient estimates within the 50% bandwidth, which contains 41 schools in 2016 (compared with 87 in the 100% CCT bandwidth), suggest the negative effects of NCT occurred in 2016 in these 41 schools closest to the cutoff. We see similar patterns when we estimate within alternative bandwidths and using the full sample; models estimated within narrower bandwidths calculated using the bandwidth selection procedure described in Imbens & Kalyanaraman (2009) show significant negative effects, while models estimated on the full sample find null effects in 2016, as shown in Table A-1. [8]

Table 5 ABOUT HERE

Central to the validity of our estimates is the ability to rule out a weak instrument (Stock & Yogo, 2002). The recommended minimum first-stage *t*-statistic on the treatment indicator to show that the instrument is a sufficiently strong predictor of treatment is 4 (What Works Clearinghouse, 2017). The first-stage *t*-statistics toward the bottom of Table 5 all exceed this criterion.

The results are qualitatively similar across subject areas, with consistently negative point estimates for math, reading, and science across all specifications in both years. The significant negative effects in 2017 appear to be driven by reading scores, where we estimate an effect of -0.16 standard deviations of test score growth for students in treatment schools (Table A-2). We also find qualitatively similar results when we estimate on test score levels rather than growth, shown in Table A-3, providing some evidence that the negative effects aren't driven by idiosyncrasies of the sample of students with lagged scores or the variation in timing for lagged score in high school and science exams. Finally, the negative effects of NCT appear to be consistent across all school levels, although we do not have a strong enough first stage to obtain valid estimates in elementary schools (Table A-4).

Figure 4 ABOUT HERE

Teacher turnover. We also find evidence that teachers in NCT schools were more likely to turn over in 2017, shown visually in Figure 4 and statistically in Table 6. Specifically, treatment school teachers were about 22.5 percentage points more likely to turn over than control school teachers in 2017. These estimates are consistent across bandwidths and robust to the inclusion of covariates. We also find that teacher turnover is significantly higher in NCT schools across the full sample (Table A-5). While a weak first stage in the 50 percent bandwidth precludes valid inferences in the fuzzy model within this bandwidth, a sharp specification finds

significant increases in teacher turnover in the narrowest bandwidth and across other bandwidths (Table A-6). Meanwhile, while teachers were descriptively less likely to turn over in NCT schools in 2016, the difference is not statistically significant, as we show in Table 6.

Table 6 ABOUT HERE

Compositional effects of teacher turnover. While teacher turnover has been found to generally have negative effects (Hanushek, Rivkin, & Schiman, 2016; Henry & Redding, 2018; Ronfeldt, Loeb, & Wyckoff, 2013), strategically replacing lower performing teachers with more effective teachers can have positive effects—especially in very low performing schools (Adnot, Dee, Katz, & Wyckoff, 2017; Henry et al., 2015; Strunk et al., 2016; Zimmer et al., 2017). By extension, a negative compositional effect of teacher turnover may help to explain negative effects on student achievement. If turnover of effective teachers was particularly high in 2016, or if replacement teachers in 2017 were worse on average than departing teachers, these staffing changes could help explain the negative effects in 2017. Meanwhile, lower turnover of effective teachers or higher turnover of ineffective teachers in 2017 might suggest that schools are engaging in strategic staffing for the future and that the negative effects in 2017 may be temporary.

We do not find consistent evidence for strategic staffing in either year. If the negative effects in 2017 were driven by turnover of more effective teachers paired with low-effectiveness replacement teachers, Table 7 would show positive point estimates on both *TOT x high effectiveness* in Panel A for 2016 (Columns 1-3) and *TOT x low effectiveness* in Panel B for 2017 (Columns 4-6). The former would suggest that highly effective teachers in treatment schools were more likely to turn over than their counterparts in control schools after the first year of services, while the latter would suggest that treatment schools were more likely than control

schools to fill vacancies with less effective teachers. We do not detect significant effects on any of these coefficients. Similarly, the estimates on *TOT x high effectiveness* in Panel A for 2016 suggest highly effective teachers were no less likely to turn over in treatment than in control schools. To that end, we do not find evidence that the negative effects were driven by a negative compositional effect of turnover.

Table 7 ABOUT HERE

Meanwhile, if the high turnover in 2017 were strategic, with treatment schools intentionally dismissing or coaching out their least effective teachers, we would observe positive estimates on *TOT x low effectiveness* in Panel A for 2017 (Columns 4-6). Significant positive effects for this group would provide evidence that the least effective teachers were more likely to turn over than their counterparts in control schools, suggesting the negative effects might be temporary as the reform schools re-staff. We do find that these estimates are descriptively positive and significant on one measure, but we also see that treatment school teachers in all three effectiveness categories were descriptively more likely to turn over in 2017 than their counterparts in control schools. Taken together, these findings suggest teacher mobility in treatment schools was neither detrimental enough in 2016 to explain student achievement losses, nor was it clearly strategic in 2017 to augur future growth. Still, we cannot completely rule out either of these hypotheses given the relatively imprecise estimates in some of these models.

Validity Checks

Two assumptions are critical to the validity of the RD design. First, there should be no manipulation of the forcing variable or cutoff; in other words, there should be no evidence that the value of the performance composite or the eligibility threshold was changed to influence treatment assignment in schools near the cutoff. Second, the functional form of the relationship

between the outcome and forcing variable must be correctly specified on both sides of the cutoff. Additional essential assumptions for the validity of the fuzzy RD design are that treatment eligibility is a sufficiently strong predictor of compliance with assignment to treatment and there is no clear violation of the exclusion restriction. In this section, we describe the above assumptions in detail and then provide evidence that the data meet additional assumptions relevant to the validity and consistency of our estimates.

As described in the Data section above, the state determined the cutoff value of the assignment variable after schools administered exams based on the number of schools that could be served by NCT. Manipulation by schools is therefore highly unlikely because schools did not know before the exam window the proficiency rate threshold for assignment to treatment. Even so, we demonstrate the integrity of the forcing variable graphically and statistically. Figure 5 shows the density of the forcing variable for the full sample of eligible schools, with a vertical line at 0 denoting the cutoff. While the histogram shows a small jump in density just above the cutoff, the jump does not represent a significant discontinuity in the density forcing variable. A McCrary test fails to reject the null of that there is no discontinuity in the density of the forcing variable within the optimal CCT 2016 and 2017 bandwidths.[9]

Figure 5 ABOUT HERE

The second core assumption for the validity of the local average treatment effect estimate is that the functional form is correctly specified on either side of the forcing variable. To meet this condition, we estimate separate local linear regressions within the CCT bandwidths on either side of the cutoff. Figure 2 and Figure 4 above provide visual evidence that the relationships are linear within the preferred bandwidths for student achievement and teacher turnover, respectively. We also estimate effects within several alternative bandwidths, including 50

percent of the CCT bandwidth, 200 percent of the CCT bandwidth, the IK optimal bandwidth, and 200 percent of the IK bandwidth, and find that both outcomes are robust to most of these alternative bandwidths and on the full sample (Table A-1 and Table A-5).

The fuzzy RD design requires that eligibility is a sufficiently strong predictor of participation. Figure 3 above clearly shows schools below the cutoff had a high probability of receiving services while schools below the cutoff had a low probability of receiving treatment. First-stage test statistics on the treatment eligibility indicator provide formal evidence that the forcing variable is a sufficiently strong predictor of participation. All first-stage test statistics on the treatment indicator are above the minimum recommended threshold of 4 (What Works Clearinghouse, 2017) in our preferred models as described in the Results section above. The first stage does not meet suggested criteria for narrower alternative bandwidths in the teacher turnover models or for the elementary school models. We denote models with weak first stages using a red box around the test statistic.

As we describe in the Results section above, we find no evidence that the instrument fails to meet the exclusion restriction; that is, whether a school scores above or below the cutoff does not appear to affect either of the outcomes through a channel other than the treatment itself. While we cannot directly test this assumption, the similarity of the ITT and TOT estimates shown in Table 4 above for student achievement growth in Table A-6 for teacher turnover support the validity of the TOT estimates.

Another key assumption for the RD estimates to be consistent is that relationship between the forcing variable and outcome would be smooth in the absence of the intervention. While we cannot test this condition directly because we cannot observe the outcomes for treatment schools in the absence of treatment, we provide evidence for the smoothness condition in two ways.

First, we show that the treatment and control samples are balanced on several key variables associated with school performance, conditional on the forcing variable, within the 2016 and 2017 preferred CCT bandwidths. Table 8 shows results from a series of models estimating the baseline covariate value using the forcing variable and a triangular kernel within the preferred bandwidths. None of the *p*-values indicate statistical significance, which demonstrates the treatment and control samples are balanced on observed covariates within our preferred bandwidths—providing evidence that assignment to treatment approximates random assignment in the region around the cutoff.

Table 8 ABOUT HERE

Graphical analysis provides further evidence that the data meet the smoothness condition (Figure 2 and Figure 4), and we conduct an additional test in which we specify a series of placebo cutoffs and test for discontinuities. We find no evidence of significant discontinuities across multiple placebo cutoffs above and below the threshold in 2016 or 2017 (Table A-7).

As a final check, we test for differential attrition across the treatment and control schools. Three schools closed during the study period—one control and two treatment schools. Of those three schools, one treatment and one control school are within the optimal CCT bandwidth for both 2016 and 2017. The overall and differential levels of attrition both fall below the conservative boundary set in the What Works Clearinghouse standards (What Works Clearinghouse, 2017).

Table 9 ABOUT HERE

Explaining the negative effects

While the evidence we presented supports the validity of the negative effects of NCT, these negative effects are puzzling given the benign nature of the intervention. To attempt to

explain these negative effects of NCT on student achievement and teacher retention, we examined the moderating influence of three elements of implementation. First, we developed a measure of fidelity of implementation (FOI) based on the theory of change and tested for heterogeneity by FOI level. We hypothesized that treatment schools that experienced low levels of implementation fidelity may have more negative effects because the services they received did not match the expected treatment. Second, we used coaching reports provided by DPI to construct dosage measures for instructional coaching and school transformation coaching, and examined whether dosage level moderated treatment effects. In this case, we hypothesized that schools receiving a lower dosage of services may have experienced negative effects if the amount of coaching received fell short of expectations and frustrated principals and teachers. Finally, we tested for heterogeneous effects based on CNA timing. Because all services were intended to build from the CNA, we hypothesized that schools not receiving CNAs or not receiving them within a useful time period for planning might suffer from less coherent services that would exacerbate already existing barriers to improvement or interrupt ongoing reform strategies.

To test for these heterogeneous effects, we run fuzzy RD models with multiple treatments (high, mid, and low dosage; high, mid, and low FOI; and by CNA timing) within our preferred bandwidths. We do not find that fidelity of implementation or dosage had moderating effects on student achievement (see Figure A-1 and Figure A-2). However, we find evidence that CNA timing is associated with variation of test score gains. In particular, Figure 6 shows that the negative effects in 2017 were concentrated in schools that did not receive CNAs at all, received CNAs in 2014 or 2015 before the intervention began, or received CNAs in spring 2016 while school improvement plans were being implemented and coaching services were rolling out. We

observe null effects in the 17 schools that received CNAs in the 2016-17 school year, a period in which the findings of the CNAs could have been integrated with revisions to the school improvement plans.

Figure 6 ABOUT HERE

Qualitative data collected as a part of the overall evaluation provides some context for interpreting these results. Descriptively, the strongest negative effects appear in schools that never received a CNA or received a CNA more than two years before the NCT services began. The intervention delivered in these schools effectively undermined the theory of change, which predicated the reform strategy on an in-depth needs assessment drawing from multiple forms of data, including instructional observations. Schools receiving CNAs in 2014 or 2015, prior to the implementation of NCT in 2016, also present negative effects. These schools received services based on findings from before they were designated as eligible for NCT and, in many schools, before much of the staff carrying out the school improvement plans, including the principals, were in place. To that end, the needs identified among these schools—such as instructional quality in specific subjects or grades that are observed as part of the CNA process—may have been outdated, and services aligned to these needs again may have been misaligned with the current needs of the school. Moreover, the principals and school improvement teams in these schools may have been unaware that the CNA was conducted or of the particular needs that were identified, and thus unable to take the findings into account during the school improvement planning.

Finally, schools that received CNAs in spring 2016, which experienced negative effects that were descriptively weaker than the latter two groups, may have struggled due to two factors. First, CNA findings communicated in the middle of the school year may have disrupted

implementation of the school improvement plan that was prepared during the prior fall, undermining commitment to the plan when school staff were preparing for state testing. Second, data collected from teachers and principals in the schools receiving CNAs in spring 2016 suggested weak communication between state and school staff concerning the CNA timing and process. During this time period, state agency personnel communicated about the CNA with principals and expected principals to communicate with their staff. Principals and teachers in these schools shared that they felt intimidated by state personnel conducting the CNAs, many staff took offense when observers showed up in their classrooms without prior notice, and many were demoralized by the description of the schools' inadequacies presented in the CNA reports after they had committed substantial effort to implementing the improvement plan. The evaluation team shared these formative findings with NCT leadership and staff in summer 2016, and later qualitative data collection suggests program staff became much more proactive in their communication with the schools receiving CNAs, which corrected the communication issues that arose in spring 2016. School staff perceived CNAs conducted during the 2016-17 school year more favorably, and our findings show no negative effects among this group of schools.

Discussion

We find that NCT had a negative effect on student achievement growth and teacher retention. The negative effects on student achievement growth appeared to begin in 2016 in the schools closest to the cutoff and then extended to the full sample in 2017. The increase in teacher turnover occurred in the second year of the intervention. Our results are robust to alternative bandwidths, ITT (sharp) and TOT (fuzzy) specifications, and to the inclusion of covariates. Because NCT aligns with ESSA requirements for school turnaround and—like turnaround under ESSA—was funded through a reallocation of existing funds rather than an infusion of new

federal funds, these negative effects serve as a cautionary tale for states developing their plans to serve low-performing schools. States are charged with continuously improving their lowest performing schools but no are longer required to take actions to disrupt the status quo and no longer receive substantial additional funds.

While the increased teacher turnover in NCT schools in 2017 opens the possibility of strategic staffing by replacing less effective teachers with more effective ones, our findings do not support this possibility. While ineffective teachers were descriptively more likely to turn over in treatment schools than control schools, average and effective teachers followed similar patterns. In other words, treatment schools experienced higher turnover across the range of teacher effectiveness, with ineffective teachers no more likely to turn over than more effective teachers in their schools. This finding suggests teacher turnover was likely voluntary on the part of teachers rather than the result of strategic staffing. This turnover did not serve to disrupt the status quo in these schools because it is already the status quo; in the year prior to treatment about 30 percent of teachers turned over in treatment schools. In turnaround models that increased student achievement, such as in Massachusetts, Tennessee (iZones), and Los Angeles, researchers found evidence that staffing decisions were strategic, with deselection of the lowest performing teachers, intentional recruitment and retention of higher performing teachers, and professional development and support for all teachers (Papay & Hannon, 2018; Strunk et al., 2016; Zimmer et al., 2017).

In North Carolina, DPI provided coaching support for teachers and principals, but the amount of coaching varied across and within schools. Rather than building school capacity through strategic staffing and comprehensive professional development, NCT focused on coaching to develop individual teacher skills and capacity in schools where, on average, the

entire staff turns over every three years. Also, strategic staffing is likely a less viable strategy in this largely rural sample of schools than in urban or suburban schools that can draw from a larger pool of educators in the local labor market. Developing individual capacities may be an essential component of turnaround in rural schools, but our findings suggest it is not sufficient on its own—and on its face is unlikely to be an effective strategy unless complementary reforms are implemented to reduce the turnover of the teachers who have increased their instructional skills.

While school turnaround as an intervention aims to effect rapid rather than incremental change in low-performing schools (Herman et al., 2008) and some prior studies have found evidence for successful rapid turnaround (Papay & Hannon, 2018; Strunk et al., 2016; Zimmer et al., 2017), it is possible that the intended effects of school reform may not immediately translate into test score growth (see, e.g., Carlson & Lavertu, 2018) or that positive effects could grow over multiple years following the intervention (Papay & Hannon, 2018). While we cannot know for certain whether the first two years of NCT laid the groundwork for improvement in future years, we find no evidence that delayed positive effects are emerging. For example, the NCT theory of change focused largely on building the capacity of individual teachers and principals, but many of those teachers left NCT schools in 2017, taking any increased capacity with them. Additionally, because of this emphasis on individual-level capacities, it is unlikely that the intervention fostered the development of school-level systems and processes required to sustain long-term school improvement.

The results that rely on a smaller sample of schools closest to the cutoff (i.e., within the narrowest bandwidths) suggest that student achievement may have declined in the first year of the intervention in the schools around the assignment cutoff. Because NCT was a tailored intervention, the treatment schools closest to the cutoff were the highest performing of the

treatment schools. Given limited resources at DPI, which did not place instructional coaches in every treatment school and delivered services of varying intensity across schools, these were likely schools that received the lightest touch intervention. It is possible that targeting all schools in the bottom 5 percent produced negative effects in the higher achieving lowest-performing schools by spreading resources too thin. Specifically, providing limited, inconsistent supports in these schools may have contributed to an already unstable school environment. Under ESSA, states are required to designate the bottom 5 percent of schools as low performing but are not necessarily required to serve the full 5 percent with the same reform model. Larger negative effects in the higher achieving of the lowest performing schools—beginning in the first year and becoming larger in the second year—suggest states might not be able or willing to allocate sufficient resources to effectively serve all schools in the lowest 5 percent of performance. In addition, the differential effects that appear to be associated with the conduct and timing of comprehensive needs assessments—which are mandated in the ESSA legislation—point to the importance of implementation and finding the resources, both human and financial, to conduct the needs assessments prior to the school improvement planning and implementation.

Finally, the generalizability of these findings should be considered in the context of the sample. North Carolina focused its efforts on mostly rural schools. A theory of action that hindered student achievement in this sample of schools would not necessarily have the same effects in urban or suburban settings. However, 28 percent of schools in the United States are rural and 19 percent of public school students attend rural schools—with states ranging from having 8 percent of schools in rural areas (New Jersey) up to 75 percent of schools (South Dakota) (U.S. Department of Education, 2016). Low-performing schools are in rural, suburban, town, and urban contexts, and school turnaround under ESSA will target schools in each of these

contexts. Additionally, many of the lessons learned under NCT are likely applicable beyond the rural context. For example, North Carolina made decisions to spread limited resources across a large number of schools and to rely on a theory of change that does not effectively transform school-level processes and practices nor systematically address staff turnover. These strategies do not align with those implemented as part successful turnaround models in other states.

Conclusion

As states implement plans to support the lowest performing schools under ESSA, the effects of NCT on student achievement and teacher turnover suggest that school reform without intentional disruption of the status quo—particularly strategic staffing that includes a focus on both hiring and retaining more effective teachers—has the potential to hinder student achievement growth and increase unintentional teacher turnover. This analysis also suggests that direct service provision without the backing of an influx of funding—as was the case of ARRA, which allotted $4.35 billion for RttT and $3 billion for SIG—may not be a viable turnaround strategy across the entire bottom 5 percent of schools.

While these findings provide some descriptive evidence to explain the mechanisms underlying the negative effects of NCT on student achievement, future research could measure implementation fidelity and quality and changes to school morale and climate in order to examine the extent to which variation in implementation and the environment mediate or suppress the effects of interventions to improve the lowest performing schools.

---

[1] While states may choose to follow school reform models that parallel the four RttT/SIG models, a separate analysis of all state ESSA plans shows very few states have committed to doing so. A total of five states outlined policies in their ESSA plans that committed to state takeover, transferring low-performing schools to alternative management, or staff replacement.

[2] About 26% of teachers with lagged EVAAS scores are low EVAAS, 63% are mid, and 11% are high.

[3] On NCEES standard 1, about 49% of teachers with lagged scores in the sample are high, 49% are mid, and 2% are low. On standard 4, about 41% are high, 57% are mid, and 2% are low.

[4] Because lagged test scores vary by subject area and grade level, we also estimate models without the lagged test score and find similar results.

[5] We do not estimate on 50% of the IK bandwidth because the bandwidth size—which unlike the CCT procedure does not account for the clustering of students within schools—includes only three schools above the cutoff.

[6] We also estimate the same set of test score models clustering standard errors at the student level to account for clustering of students across multiple exams in a year. However, the standard errors clustered at the student level are smaller, so the estimates with standard errors clustered at the school level that we show represent a more conservative approach.

[7] Because we include the lagged test score on the right side of the model, the estimated effect on student achievement growth in 2017 represents the effect of NCT in the second year of services after partialing out any effect from the first year.

[8] The IK bandwidths are narrower than the CCT bandwidths. The estimate in the 17 schools within the IK bandwidth is -.186 and the estimate in the 35 schools in the 200% IK bandwidth is -.146.

[9] 2016 p=.2768; 2017 p=.1773

References

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

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*(434), 444–455. https://doi.org/10.1080/01621459.1996.10476902

Bonilla, S., & Dee, T. (2017). *The Effects of School Reform Under NCLB Waivers: Evidence from Focus Schools in Kentucky* (Working Paper No. 23462). https://doi.org/10.3386/w23462

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

Carlson, D., & Lavertu, S. (2018). School Improvement Grants in Ohio: Effects on Student Achievement and School Administration. *Educational Evaluation and Policy Analysis*, 0162373718760218. https://doi.org/10.3102/0162373718760218

Dee, T. (2012). *School Turnarounds: Evidence from the 2009 Stimulus* (Working Paper No. 17990). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w17990

Dee, T., & Dizon-Ross, E. (2019). School Performance, Accountability, and Waiver Reforms: Evidence From Louisiana. *Educational Evaluation and Policy Analysis*, 0162373719849944. https://doi.org/10.3102/0162373719849944

Dickey-Griffith, D. (2013). Preliminary effects of the school improvement grant program on student achievement in Texas. *The Georgetown Public Policy Review*, 21–39.

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

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, C., … Giffin, J. (2017). *School Improvement Grants: Implementation and Effectiveness. NCEE 2017-4013*. Retrieved from https://eric.ed.gov/?id=ED572215

Gross, B., Booker, T. K., & Goldhaber, D. (2009). Boosting Student Achievement: The Effect of Comprehensive School Reform on Student Achievement. *Educational Evaluation and Policy Analysis*, *31*(2), 111–126. https://doi.org/10.3102/0162373709333886

Hanushek, E. A., Rivkin, S. G., & Schiman, J. C. (2016). Dynamic effects of teacher turnover on the quality of instruction. *Economics of Education Review*, *55*, 132–148. https://doi.org/10.1016/j.econedurev.2016.08.004

Heissel, J. A., & Ladd, H. F. (2018). School turnaround in North Carolina: A regression discontinuity analysis. *Economics of Education Review*, *62*, 302–320. https://doi.org/10.1016/j.econedurev.2017.08.001

Hemelt, S. W., & Jacob, B. (2017). *Differentiated Accountability and Education Production: Evidence from NCLB Waivers* (Working Paper No. 23461). https://doi.org/10.3386/w23461

Hemelt, S. W., & Jacob, B. A. (2018). How Does an Accountability Program that Targets Achievement Gaps Affect Student Performance? *Education Finance and Policy*, 1–68. https://doi.org/10.1162/edfp_a_00276

Henry, G. T., Campbell, S. L., Thompson, C. L., & Townsend, L. W. (2014). *Evaluation of District and School Transformation School-Level Coaching and Professional Development Activities*.

Henry, G. T., Guthrie, J. E., & Townsend, L. W. (2015). *Outcomes and Impacts of North Carolina's Initiative to Turn Around the Lowest-Achieving Schools*. Retrieved from http://cerenc.org/wp-content/uploads/2015/09/ES-FINAL-Final-DST-Report-9-3-15.pdf

Henry, G. T., & Redding, C. (2018). The consequences of leaving school early: The effects of within-year and end-of-year teacher turnover. *Education Finance and Policy*, 1–52.

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., & Redding, S. (2008). *Turning Around Chronically Low-Performing Schools: A Practice Guide* (No. NCEE 2008-4020). Retrieved from National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education website: http://ies.ed.gov/ncee/wwc/PracticeGuide.aspx?sid=7

Imbens, G., & Kalyanaraman, K. (2009). *Optimal Bandwidth Choice for the Regression Discontinuity Estimator* (Working Paper No. 14726). https://doi.org/10.3386/w14726

Imbens, G., & Lemieux, T. (2007). *Regression Discontinuity Designs: A Guide to Practice* (Working Paper No. 13039). https://doi.org/10.3386/w13039

Papay, J., & Hannon, M. (2018, November 8). *The Effects of School Turnaround Strategies in Massachusetts*. Presented at the 2018 APPAM Fall Research Conference: *Evidence for Action: Encouraging Innovation and Improvement*, Washington, D.C. Retrieved from https://appam.confex.com/appam/2018/webprogram/Paper26237.html

Peurach, D., & Neumerski, C. (2015). Mixing metaphors: Building infrastructure for large scale school turnaround. *Journal of Educational Change*, *16*(4), 379–420. https://doi.org/10.1007/s10833-015-9259-z

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How Teacher Turnover Harms Student Achievement. *American Educational Research Journal*, *50*(1), 4–36. https://doi.org/10.3102/0002831212463813

Schueler, B. E., Goodman, J., & Deming, D. J. (2016). *Can States Take Over and Turn Around School Districts? Evidence from Lawrence, Massachusetts* (Working Paper No. 21895). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w21895

Stock, J. H., & Yogo, M. (2002). *Testing for Weak Instruments in Linear IV Regression* (Working Paper No. 284). https://doi.org/10.3386/t0284

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_a_00188

Sun, M., Penner, E. K., & Loeb, S. (2017). Resource- and Approach-Driven Multidimensional Change: Three-Year Effects of School Improvement Grants. *American Educational Research Journal*, *54*(4), 607–643. https://doi.org/10.3102/0002831217695790

Thompson, C. L., Brown, K. M., Townsend, L. W., Henry, G. T., & Fortner, C. K. (2011). Turning around North Carolina's lowest achieving schools (2006-2010). *Consortium for Educational Research and Evaluation–North Carolina*.

U.S. Department of Education. (2016). *Public Elementary/Secondary School Universe Survey* (No. Provisional Version 1a, and the NCES Education Demographic and Geographic Estimates (EDGE)). National Center for Education Statistics (NCES), Common Core of Data (CCD).

What Works Clearinghouse. (2017). *Standards Handbook. Version 4.0.* Retrieved from U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. website: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

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. CNA and unpacking timing

	Number of schools	
	CNA	Unpacking
2014-2015	17	15
Spring 2016	25	4
Summer 2016	0	24
Fall 2016	15	1
Spring 2017	4	2
Summer 2017	0	3
Fall 2017	2	0
Pending	0	4
None during intervention period	12[a]	22[b]
Total schools: 75		

[a] Of these 12 schools that did not receive CNAs, four declined and eight were not conducted due to Hurricane Matthew.

[b] Of these 22 schools that did not receive unpackings, 12 were schools without CNAs, two declined, two were schools that had received CNAs in fall 2017 because they were under consideration for the state's Innovative School District (ISD), and the remaining six were not conducted for unknown reasons.

Table 2. Coaching visits

	Instructional	School transformation
Total schools with coaches assigned	65 total 16 math, 18 ELA, 12 science, 33 non-subject-specific	56 total
Number of visits	Range: 0-137 Mean: 45.36	Range: 0-63 Mean: 25.28
Visits per teacher	Range: 0-15.75 Mean: 1.83	Range: 0-3.82 Mean: 1.03

Source: DPI coaching reports.

NOTE: Subject-level and non-subject-specific ICs do not add up to 65 because schools have ICs focused on multiple subjects. Means are for all treatment schools regardless of whether they have a coach assigned. Visits per teacher based on number of FTE teachers employed in the school across all treatment schools.

Table 3. School sample characteristics

	NCT	Control
Urbanicity		
City	0.0 (0.11)	0.0 (0.16)
Suburb	0.0 (0.00)	0.0 (0.19)
Town	0.1 (0.25)	0.1 (0.28)
Rural	0.9 (0.27)	0.8 (0.36)
School level		
Elementary	48.7 (50.31)	50.0 (50.32)
Middle	35.9 (48.28)	35.0 (48.00)
High	15.4 (36.31)	15.0 (35.93)
Student achievement		
2015 performance composite (centered)	-5.1 (4.31)	3.5 (1.77)
EVAAS growth score	68.6 (10.36)	68.5 (11.31)
Teacher qualifications		
Percent novice teachers	32.5 (12.57)	27.6 (11.66)
Percent National Board Certification	7.7 (4.64)	10.3 (6.14)
Student demographics		
Minority percent	84.7 (12.44)	72.3 (15.68)
Economically disadvantaged	82.2 (12.12)	77.8 (13.95)
School characteristics		
Per pupil spending	10217.70 (2264.00)	9698.90 (1602.07)
Average Daily Membership	429.2 (172.67)	479.6 (223.59)

NOTE: Means and standard deviations on baseline measures based on 78 treatment and 80 control schools.

Table 4. ITT estimates *(outcome=test score growth)*

Panel A: 2016

	(1) CCT	(2)	(3) 50% CCT	(4)	(5) 200% CCT	(6)
ITT	-0.063	-0.034	-0.130*	-0.090**	-0.024	-0.020
	(0.0581)	(0.0443)	(0.0583)	(0.0295)	(0.0403)	(0.0356)
Covariates		X		X		X
Bandwidth	4.1	4.1	2.1	2.1	8.3	8.3
N	195437	195437	195437	195437	195437	195437
N Bandwidth	50731	50731	23415	23415	92514	92514
T schools in BW	36	36	22	22	66	66
C schools in BW	51	51	19	19	102	102

Panel B: 2017

	(1) CCT	(2)	(3) 50% CCT	(4)	(5) 200% CCT	(6)
ITT	-0.123*	-0.131**	-0.172**	-0.221***	-0.101*	-0.093*
	(0.0521)	(0.0403)	(0.0535)	(0.0249)	(0.0395)	(0.0365)
Covariates		X		X		X
Bandwidth	3.3	3.3	1.7	1.7	6.7	6.7
N	195099	195099	195099	195099	195099	195099
N Bandwidth	39423	39423	18624	18624	77420	77420
T schools in BW	31	31	18	18	55	55
C schools in BW	37	37	13	13	84	84

Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models. All models include lagged score and subject fixed effects on the right side, with math as the reference category.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5. TOT estimates *(outcome=test score growth)*

Panel A: 2016

	(1) CCT	(2)	(3) 50% CCT	(4)	(5) 200% CCT	(6)
TOT	-0.066	-0.039	-0.148**	-0.135***	-0.027	-0.023
	(0.0592)	(0.0512)	(0.0511)	(0.0389)	(0.0449)	(0.0407)
Covariates		X		X		X
Bandwidth	4.1	4.1	2.1	2.1	8.3	8.3
First-stage *t*-stat	10.99	10.66	5.94	6.07	14.60	14.61
N	195437	195437	195437	195437	195437	195437
N Bandwidth	50731	50731	23415	23415	92514	92514
T schools in BW	36	36	22	22	66	66
C schools in BW	51	51	19	19	102	102

Panel B: 2017

	(1) CCT	(2)	(3) 50% CCT	(4)	(5) 200% CCT	(6)
TOT	-0.131*	-0.170**	-0.198***	-0.420***	-0.111*	-0.109*
	(0.0517)	(0.0541)	(0.0560)	(0.0710)	(0.0433)	(0.0433)
Covariates		X		X		X
Bandwidth	3.3	3.3	1.7	1.7	6.7	6.7
First-stage *t*-stat	9.04	9.42	5.46	7.01	14.90	14.54
N	195099	195099	195099	195099	195099	195099
N Bandwidth	39423	39423	18624	18624	77420	77420
T schools in BW	31	31	18	18	55	55
C schools in BW	37	37	13	13	84	84

Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models. All models include lagged score and subject fixed effects on the right side, with math as the reference category.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6. TOT estimates *(outcome=teacher turnover)*

Panel A: 2016

	(1) CCT	(2)	(3) 50% CCT	(4)	(5) 200% CCT	(6)
TOT	-0.044	-0.103	0.091	0.164	-0.074	-0.087
	(0.0945)	(0.0822)	(0.1316)	(0.1388)	(0.0600)	(0.0547)
Covariates		X		X		X
Bandwidth	4.1	4.1	2.1	2.1	8.3	8.3
First-stage *t*-stat	5.61	5.71	2.80	2.54	8.61	8.76
N	10770	10770	10770	10770	10770	10770
N Bandwidth	2658	2658	1240	1240	5270	5270
T schools in BW	35	35	21	21	64	64
C schools in BW	51	51	19	19	102	102

Panel B: 2017

	(1) CCT	(2)	(3) 50% CCT	(4)	(5) 200% CCT	(6)
TOT	0.225**	0.204*	0.357**	0.393	0.128	0.126*
	(0.0819)	(0.0891)	(0.1342)	(0.2676)	(0.0669)	(0.0604)
Covariates		X		X		X
Bandwidth	3.3	3.3	1.7	1.7	6.7	6.7
First-stage *t*-stat	4.90	4.98	2.62	2.35	8.17	8.31
N	10492	10492	10492	10492	10492	10492
N Bandwidth	2078	2078	940	940	4280	4280
T schools in BW	30	30	17	17	53	53
C schools in BW	37	37	13	13	84	84

Estimates from linear probability models. Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models. Red outlines denote first-stage test statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7. TOT estimates on teacher turnover and new-to-school teachers by lagged teacher effectiveness

Panel A: Teacher turnover

	2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)
	Standard 1	Standard 4	EVAAS	Standard 1	Standard 4	EVAAS
Low effectiveness	-0.092	0.211	0.078	-0.338	0.083	0.060
	(0.1278)	(0.1309)	(0.0615)	(0.4532)	(0.1457)	(0.0533)
High effectiveness	-0.048	-0.036	-0.088	-0.029	0.002	-0.022
	(0.0376)	(0.0540)	(0.0742)	(0.0231)	(0.0281)	(0.1021)
TOT x low effectiveness	0.082	-0.205	-0.142	1.023	0.558*	0.157
	(0.1823)	(0.1989)	(0.1215)	(0.6286)	(0.2722)	(0.1185)
TOT x mid effectiveness	-0.057	-0.045	-0.017	0.221*	0.193*	0.137
	(0.0911)	(0.0880)	(0.1242)	(0.0952)	(0.0879)	(0.1059)
TOT x high effectiveness	0.025	0.055	0.004	0.155	0.186	0.104
	(0.1142)	(0.1338)	(0.1803)	(0.0820)	(0.1006)	(0.1296)
Constant	0.295***	0.274***	0.261**	0.166***	0.154***	0.188**
	(0.0841)	(0.0825)	(0.1003)	(0.0395)	(0.0382)	(0.0594)
N	1997	1997	1102	1568	1568	786

Panel B: New-to-school teachers

	2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)
	Standard 1	Standard 4	EVAAS	Standard 1	Standard 4	EVAAS
Low effectiveness	0.241*	-0.067***	0.032	0.307	0.069	-0.007
	(0.1223)	(0.0124)	(0.0383)	(0.2315)	(0.0915)	(0.0431)
High effectiveness	-0.036**	-0.054***	0.057	0.005	-0.019	0.026
	(0.0122)	(0.0117)	(0.0667)	(0.0137)	(0.0140)	(0.0319)
TOT x low effectiveness	-0.150	0.148	0.021	-0.375	-0.134	0.099
	(0.1624)	(0.0837)	(0.0641)	(0.2769)	(0.1336)	(0.0535)
TOT x mid effectiveness	-0.010	-0.015	0.094	0.010	-0.008	0.098
	(0.0257)	(0.0247)	(0.0544)	(0.0168)	(0.0166)	(0.0750)
TOT x high effectiveness	0.016	0.019	0.000	-0.005	0.027	0.029
	(0.0260)	(0.0208)	(0.1295)	(0.0213)	(0.0237)	(0.0818)
Constant	0.055**	0.062***	0.043	0.022	0.031*	0.035
	(0.0198)	(0.0173)	(0.0270)	(0.0128)	(0.0134)	(0.0527)
N	1997	1997	1102	1568	1568	786

NOTE: Effectiveness based on prior year NCEES (Columns 1-2 and 4-5) and EVAAS (Columns 3 and 6). NCEES standard 1 is teacher leadership. NCEES standard 4 is facilitating student learning. Low NCEES is defined as a score of 1 or 2 on 5-point scale, mid NCEES defined as score of 3, and high NCEES defined as 4 or 5. Low EVAAS is defined as an EVAAS score of <-2, which the state categorizes as not meeting expected growth, average EVAAS is defined as a score between -2 and 2, which the state categorizes as meeting expected growth, and high EVAAS is defined as an EVAAS score of >2, which the state categorizes as exceeding expected growth.

Standard errors clustered at the school level. All models estimated within CCT bandwidths calculated using the fuzzy test score models.

All first-stage test statistics are greater than the What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage, except for the test statistic for TOT x average EVAAS in Model 6, which is 3.2.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8. Sample balance conditional on forcing variable within optimal bandwidths

	2016 Treat	Control	p-value	2017 Treat	Control	p-value
School-level student demographics						
ED percent	78.18	85.37	0.262	78.31	86.00	0.254
Minority percent	75.74	74.69	0.901	75.73	73.87	0.839
Black percent	48.98	47.81	0.919	49.12	48.49	0.960
Hispanic percent	15.40	21.84	0.317	15.27	21.27	0.374
ADM	437.06	455.30	0.871	438.40	450.32	0.921
Teachers						
Novice teacher rate	0.34	0.37	0.592	0.34	0.38	0.496
Mean teaching experience	10.26	10.22	0.976	10.28	10.18	0.947
Teacher turnover	0.33	0.33	0.983	0.34	0.34	0.975
School performance						
School EVAAS	69.57	66.96	0.631	0.34	0.38	0.496

NOTE: Estimates from RD with covariate listed in row as outcome and triangular kernel. Treatment and control samples within optimal CCT bandwidths.

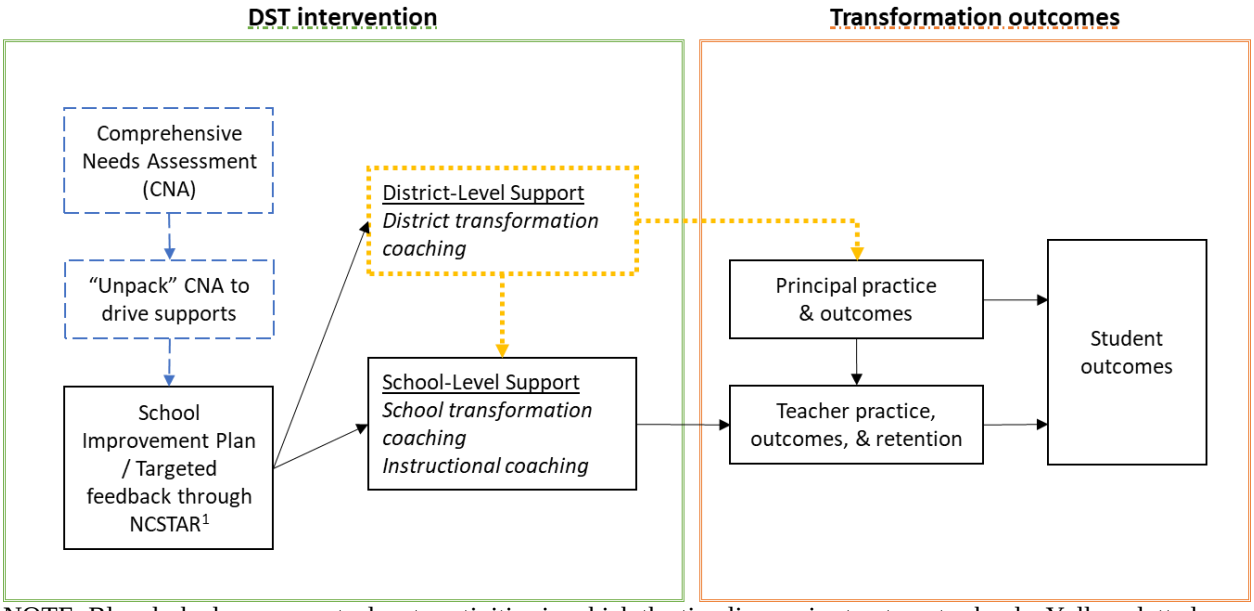
Table 9. Attrition

	CCT 2016 (4.13)	CCT 2017 (3.35)
β_{treat}	.042	.046
$\beta_{compare}$	.044	.048
$\beta_{overall}$	.043	.047
β_{diff} (SE)	-.002 (.060)	-.003 (.066)

NOTE: Estimates from linear probability model predicting attrition at the school level and controlling for the forcing variable within the optimal CCT bandwidths and with a triangular kernel.

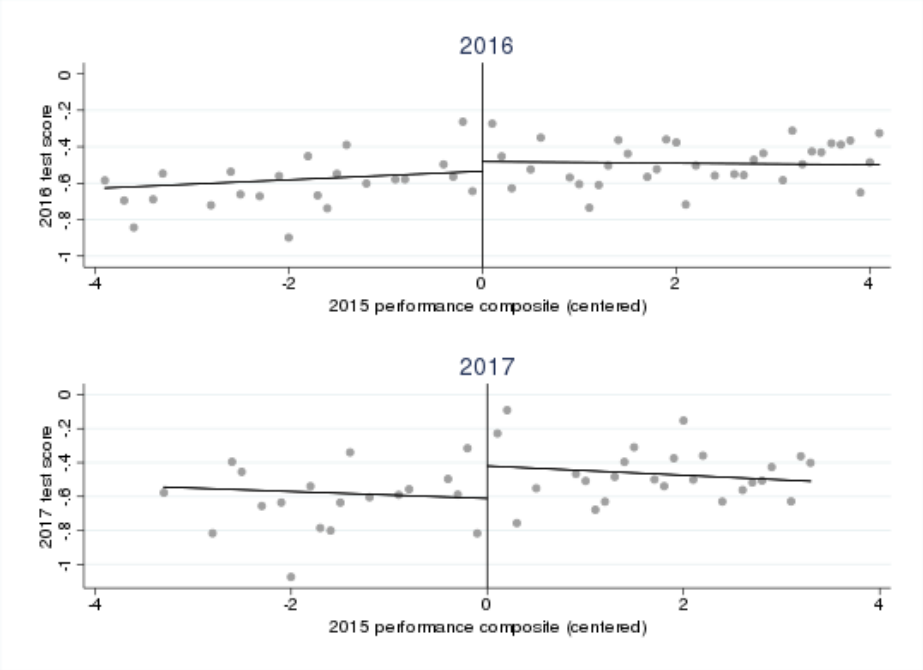
Figures

Figure 1. North Carolina Transformation Theory of Change



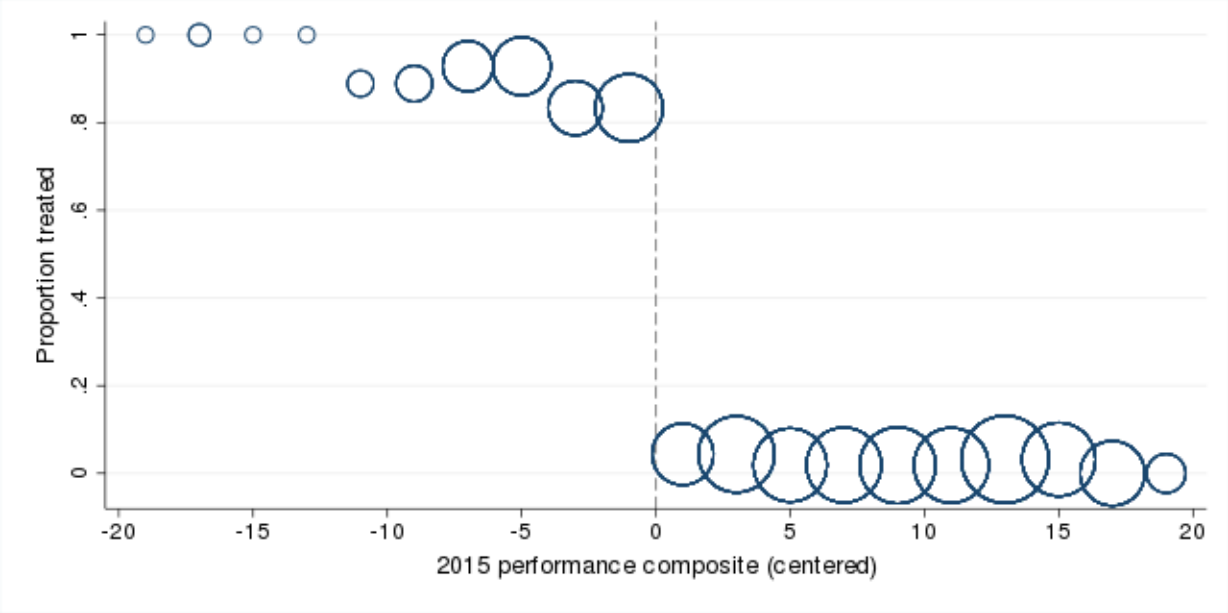
NOTE: Blue dashed components denote activities in which the timeline varies treatment schools. Yellow dotted components denotes activities not available to all districts.

Figure 2. Student achievement by distance from assignment threshold



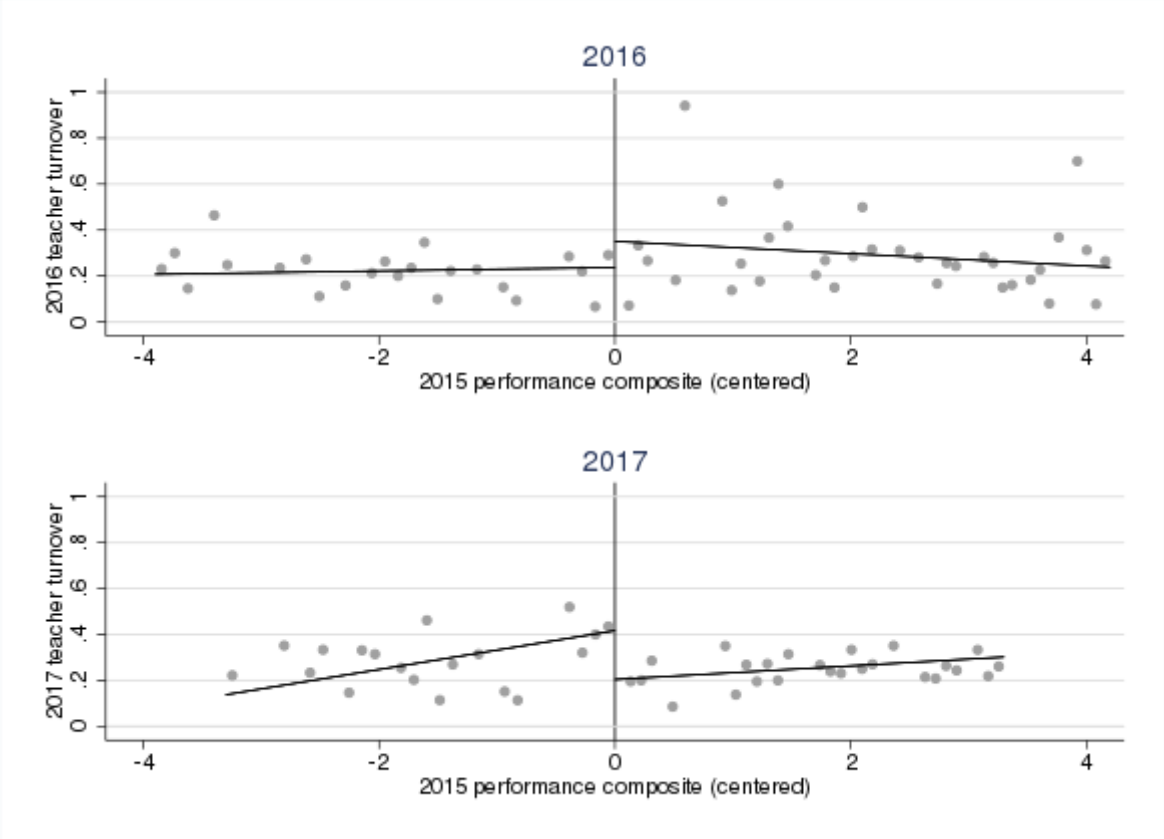
NOTE: Markers represent bin averages within CCT bandwidths and lines are linear fit. Estimation using triangular kernel within preferred CCT bandwidth, with average bin width of .006 to left of cutoff and .007 to right of cutoff in 2016, and .007 to left of cutoff and .010 to right of cutoff in 2017.

Figure 3. Proportion treated by forcing variable



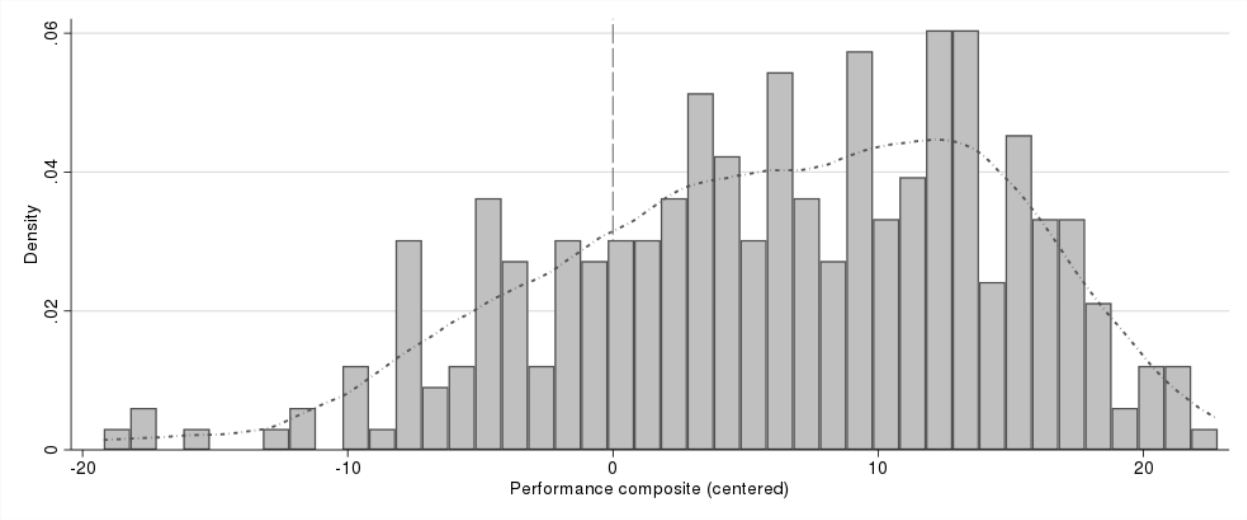
NOTE: Markers represent bin averages. Bin width is 2. Marker sizes weighted by number of schools in bin.

Figure 4. Teacher turnover by distance from assignment threshold



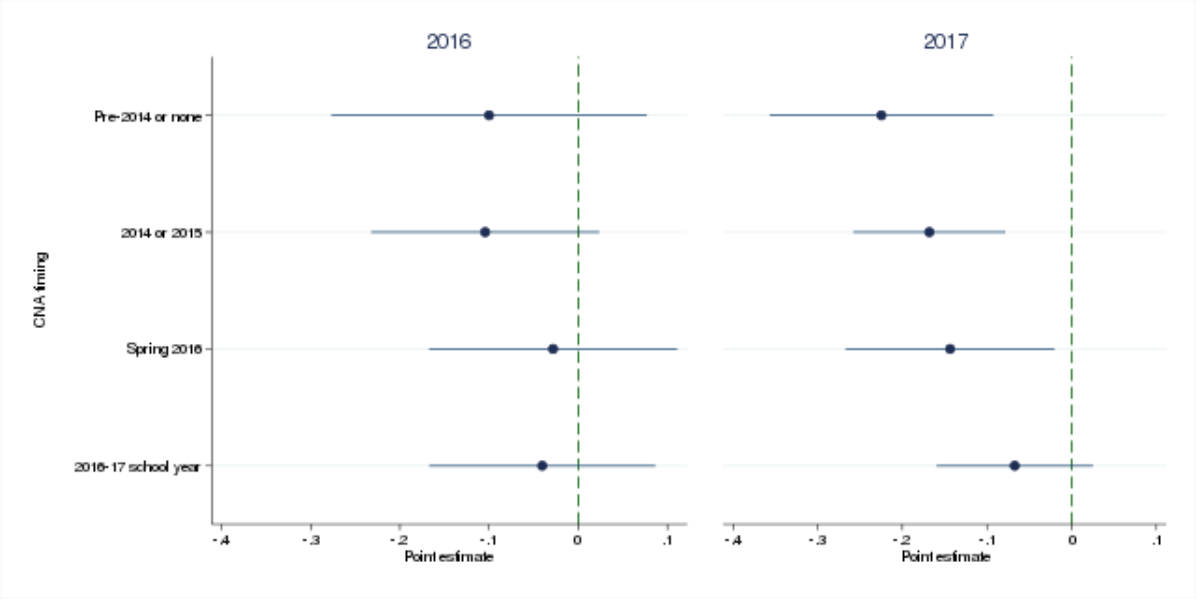
NOTE: Graph based on school-level averages of dichotomous teacher turnover variable. Markers represent individual school averages and lines are linear fit. Estimation using triangular kernel within preferred CCT bandwidth.

Figure 5. Graphical integrity of the forcing variable



NOTE: Bin width is 1. Includes all eligible schools.

Figure 6. Heterogeneity of Effects by Comprehensive Needs Assessment Timing



NOTE: Estimates from fuzzy RD models with triangular kernel and 4 different treatments within preferred CCT bandwidths. Markers represent point estimates and spikes represent 95% confidence intervals. CCT bandwidths calculated using main fuzzy test score models. All first-stage test statistics are greater than the What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage. Corresponding point estimates provided in Table A-8.

Appendix

Table A-1. TOT estimates within alternative bandwidths and full sample *(outcome=test score growth)*

Panel A: 2016

	(1) No BW	(2)	(3) IK	(4)	(5) 200% IK	(6)
TOT	-0.027	-0.017	-0.186***	1.095	-0.146**	-0.086*
	(0.0478)	(0.0437)	(0.0526)	(0.5948)	(0.0527)	(0.0381)
Covariates		X		X		X
Bandwidth	23.0	23.0	0.9	0.9	1.7	1.7
First-stage *t*-stat	12.66	12.50	8.77	1.92	5.62	5.54
N	83896	83896	195437	195437	195437	195437
N Bandwidth	83896	83896	10184	10184	20909	20909
T schools in BW	78	78	12	12	20	20
C schools in BW	80	80	5	5	15	15

Panel B: 2017

	(1) No BW	(2)	(3) IK	(4)	(5) 200% IK	(6)
TOT	-0.110**	-0.088*	-0.307***	0.042*	-0.207***	-0.413***
	(0.0417)	(0.0427)	(0.0823)	(0.0173)	(0.0606)	(0.0716)
Covariates		X		X		X
Bandwidth	23.0	23.0	0.7	0.7	1.5	1.5
First-stage *t*-stat	12.60	12.01	6.99	72.56	5.44	7.93
N	83393	83393	195099	195099	195099	195099
N Bandwidth	83393	83393	8473	8473	16740	16740
T schools in BW	78	78	11	11	15	15
C schools in BW	79	79	4	4	12	12

NOTE: Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models. All models include lagged score and subject fixed effects on the right side, with math as the reference category. 50% IK not included because the bandwidth size—which unlike the CCT procedure does not account for the clustering of students within schools—includes only three schools above the cutoff. Red outlines denote first-stage test statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-2. TOT estimates by subject *(outcome=test score growth)*

Panel A: Math

	2016			2017		
	(1) CCT	(2) 50% CCT	(3) 200% CCT	(4) CCT	(5) 50% CCT	(6) 200% CCT
TOT	-0.098	-0.141**	-0.053	-0.096	-0.159*	-0.117*
	(0.0509)	(0.0513)	(0.0489)	(0.0698)	(0.0750)	(0.0576)
Covariates	no	no	no	no	no	no
Bandwidth	4.1	2.1	8.3	3.3	1.7	6.7
First-stage *t*-stat	10.80	5.88	14.12	8.92	5.43	14.59
N	85131	85131	85131	85130	85130	85130
N Bandwidth	21766	10039	39688	17026	8086	33235
T schools in BW	36	22	66	31	18	55
C schools in BW	51	19	102	37	13	84

Panel B: Reading

	2016			2017		
	(1) CCT	(2) 50% CCT	(3) 200% CCT	(4) CCT	(5) 50% CCT	(6) 200% CCT
TOT	-0.031	-0.094	0.002	-0.164***	-0.242***	-0.129***
	(0.0567)	(0.0614)	(0.0418)	(0.0370)	(0.0517)	(0.0327)
Covariates	no	no	no	no	no	no
Bandwidth	4.1	2.1	8.3	3.3	1.7	6.7
First-stage *t*-stat	8.94	4.79	12.40	7.41	4.37	12.44
N	88535	88535	88535	88421	88421	88421
N Bandwidth	22436	10420	41286	17611	8312	34617
T schools in BW	36	22	66	31	18	55
C schools in BW	51	19	102	37	13	84

Panel C: Science

	2016			2017		
	(1) CCT	(2) 50% CCT	(3) 200% CCT	(4) CCT	(5) 50% CCT	(6) 200% CCT

TOT	-0.072	-0.326**	-0.043	-0.142	-0.187	-0.045
	(0.1658)	(0.1219)	(0.1075)	(0.1290)	(0.1491)	(0.1056)
Covariates	no	no	no	no	no	no
Bandwidth	4.1	2.1	8.3	3.3	1.7	6.7
First-stage *t*-stat	30.03	227.62	30.40	33.31	8.32e+15	38.15
N	21771	21771	21771	21548	21548	21548
N Bandwidth	6529	2956	11540	4786	2226	9568
T schools in BW	33	20	56	28	17	48
C schools in BW	50	18	97	37	13	81

NOTE: Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models. All models include lagged score on the right side. Red outlines denote first-stage test statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-3. TOT estimates without lagged test score *(outcome=test score levels)*

	2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)
	CCT	50% CCT	200% CCT	CCT	50% CCT	200% CCT
TOT	-0.054	-0.209	-0.019	-0.210	-0.429***	-0.142
	(0.1041)	(0.1120)	(0.0686)	(0.1260)	(0.1262)	(0.0872)
Covariates	no	no	no	no	no	no
Bandwidth	4.1	2.1	8.3	3.3	1.7	6.7
First-stage *t*-stat	8.53	4.49	11.90	6.92	4.00	11.51
N	235611	235611	235611	234659	234659	234659
N Bandwidth	59238	27245	109730	45948	21580	91816
T schools in BW	36	22	66	31	18	55
C schools in BW	51	19	102	37	13	84

NOTE: Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models. All models include subject fixed effects on the right side, with math as the reference category.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-4. TOT estimates by school level *(outcome=test score growth)*

Panel A: Elementary

	2016			2017		
	(1) CCT	(2) 50% CCT	(3) 200% CCT	(4) CCT	(5) 50% CCT	(6) 200% CCT
TOT	-0.025	0.039	-0.033	-0.326	-0.640	-0.293**
	(0.0785)	(0.2938)	(0.0591)	(0.1978)	(0.8842)	(0.1061)
Covariates	no	no	no	no	no	no
Bandwidth	4.1	2.1	8.3	3.3	1.7	6.7
First-stage *t*-stat	1.77	0.71	2.60	1.62	0.62	2.68
N	54933	54933	54933	56572	56572	56572
N Bandwidth	10510	4896	22309	8623	4124	20234
T schools in BW	20	10	34	16	7	29
C schools in BW	20	9	50	15	7	41

Panel B: Middle

	2016			2017		
	(1) CCT	(2) 50% CCT	(3) 200% CCT	(4) CCT	(5) 50% CCT	(6) 200% CCT
TOT	-0.069	-0.143**	-0.033	-0.090	-0.123***	-0.078
	(0.0610)	(0.0442)	(0.0546)	(0.0493)	(0.0311)	(0.0457)
Covariates	no	no	no	no	no	no
Bandwidth	4.1	2.1	8.3	3.3	1.7	6.7
First-stage *t*-stat	17.42	4.93e+15	19.54	21.03	7.20e+15	24.46
N	124063	124063	124063	122863	122863	122863
N Bandwidth	34957	16508	57805	27513	13488	48029
T schools in BW	12	9	20	12	9	16
C schools in BW	24	8	36	17	5	31

Panel C: High [a]

	2016		2017	
	(1) CCT	(2) 200% CCT	(3) CCT	(4) 200% CCT
TOT	0.022	-0.001	-0.199***	-0.112*
	(0.0428)	(0.0343)	(0.0362)	(0.0561)
Covariates				
Bandwidth	4.1	8.3	3.3	6.7
N	16441	16441	15664	15664
N Bandwidth	5264	12400	3287	9157
T schools in BW	4	12	3	10
C schools in BW	7	16	5	12

NOTE: Elementary and middle schools are estimated using fuzzy RD. High school models use a sharp RD because there is no noncompliance at the high school level.

Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models. All models include lagged score and subject on the right side, with math as the reference category. Red outlines denote first-stage test statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[a] High schools only estimated within CCT bandwidth and 200% CCT bandwidth because there are not enough high schools within the 50% bandwidth.

Table A-5. TOT estimates within alternative bandwidths and full sample *(outcome=teacher turnover)*

Panel A: 2016

	(1) No BW	(2)	(5) IK	(6)	(7) 200% IK	(8)
TOT	-0.075	-0.093	0.332*	0.078	0.152	0.331
	(0.0586)	(0.0544)	(0.1377)	(0.0792)	(0.1243)	(0.2055)
Covariates		X		X		X
Bandwidth	23.0	23.0	0.9	0.9	1.7	1.7
First-stage *t*-stat	8.86	8.77	3.47	3.32	2.47	1.96
N	4783	4783	10770	10770	10770	10770
N Bandwidth	4783	4783	488	488	1032	1032
T schools in BW	76	76	12	12	19	19
C schools in BW	80	80	5	5	15	15

Panel B: 2017

	(1) No BW	(2)	(5) IK	(6)	(7) 200% IK	(8)
TOT	0.099	0.120*	0.179*	0.056	0.378**	0.470
	(0.0511)	(0.0478)	(0.0706)	(0.0527)	(0.1453)	(0.3481)
Covariates		X		X		X
Bandwidth	23.0	23.0	0.7	0.7	1.5	1.5
First-stage *t*-stat	8.97	8.87	4.25	19.76	2.50	2.10
N	4707	4707	10492	10492	10492	10492
N Bandwidth	4707	4707	424	424	844	844
T schools in BW	76	76	11	11	15	15
C schools in BW	79	79	4	4	12	12

NOTE: Estimates from linear probability models. Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models. 50% IK not included because the bandwidth size—which unlike the CCT procedure does not account for the clustering of students within schools—includes only three schools above the cutoff. Red outlines denote first-stage test statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage. IK bandwidths calculated using the fuzzy test score models.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-6. ITT estimates *(outcome=teacher turnover)*

Panel A: 2016

	(1) CCT	(2)	(3) 50% CCT	(4)	(5) 200% CCT	(6)
ITT	-0.036	-0.075	0.066	0.068	-0.059	-0.067
	(0.0777)	(0.0610)	(0.1016)	(0.0487)	(0.0465)	(0.0411)
Covariates		X		X		X
Bandwidth	4.1	4.1	2.1	2.1	8.3	8.3
N	10770	10770	10770	10770	10770	10770
N Bandwidth	2658	2658	1240	1240	5270	5270
T schools in BW	35	35	21	21	64	64
C schools in BW	51	51	19	19	102	102

Panel B: 2017

	(1) CCT	(2)	(3) 50% CCT	(4)	(5) 200% CCT	(6)
ITT	0.187**	0.138**	0.258***	0.143**	0.104	0.096*
	(0.0668)	(0.0507)	(0.0620)	(0.0508)	(0.0554)	(0.0460)
Covariates		X		X		X
Bandwidth	3.3	3.3	1.7	1.7	6.7	6.7
N	10492	10492	10492	10492	10492	10492
N Bandwidth	2078	2078	940	940	4280	4280
T schools in BW	30	30	17	17	53	53
C schools in BW	37	37	13	13	84	84

NOTE: Estimates from linear probability models. Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-7. Placebo estimates from fuzzy RD within optimal CCT bandwidth, 2016 *(outcome=test score growth)*

Panel A: 2016

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Placebo Cutoff	-4	-3	-2	-1	1	2	3	4
TOT	-2.778	-0.200	0.058	0.058	0.171	-0.018	0.187	0.782
	(51.5073)	(1.3754)	(0.1058)	(0.1929)	(0.1863)	(0.1664)	(0.1067)	(7.6192)
Observations	195466	195466	195466	195466	195466	195466	195466	195466

Panel B: 2017

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	-4	-3	-2	-1	1	2	3	4
TOT	-0.173	1.085	-0.040	-0.136	0.019	0.004	-0.103	5.101
	(1.8885)	(14.6048)	(0.1941)	(0.1569)	(0.2192)	(0.1951)	(0.1516)	(65.7928)
Observations	195078	195078	195078	195078	195078	195078	195078	195078

NOTE: Standard errors clustered at the school level. All models include lagged score and subject on the right side, with math as the reference category. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-8. Fuzzy RD results by CNA timing *(outcome=test score growth)*

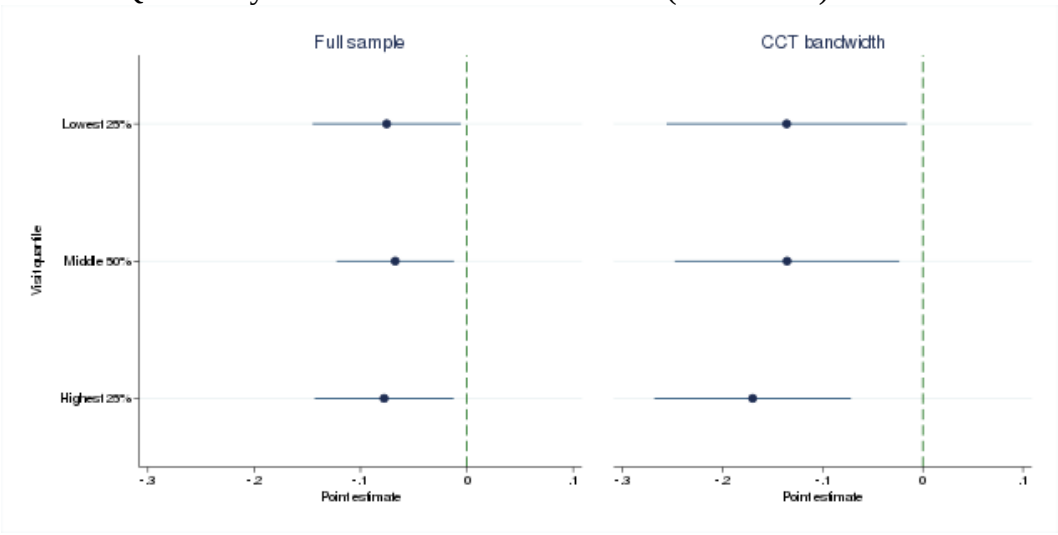
	2016		2017	
	(1)	(2)	(3)	(4)
	Full sample	CCT	Full sample	CCT
Pre-2014 or none	-0.042	-0.100	-0.203**	-0.225***
	(0.0576)	(0.0903)	(0.0754)	(0.0674)
2014 or 2015	-0.091**	-0.104	-0.087**	-0.168***
	(0.0324)	(0.0653)	(0.0333)	(0.0458)
Spring 2016	-0.027	-0.028	-0.114***	-0.144*
	(0.0360)	(0.0710)	(0.0330)	(0.0630)
2016-17 school year	-0.004	-0.040	-0.029	-0.067
	(0.0302)	(0.0647)	(0.0252)	(0.0471)
Constant	-0.103***	-0.078	-0.102***	-0.075*
	(0.0188)	(0.0450)	(0.0181)	(0.0354)
N	86354	51969	85808	39427

NOTE: 2SLS estimates from fuzzy RD using triangular kernel with four separate treatments by CNA timing. Standard errors clustered at the school level. CCT bandwidths calculated using the fuzzy test score models. All first-stage test statistics are greater than What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage. All models include lagged score and subject fixed effects on the right side, with math as the reference category.

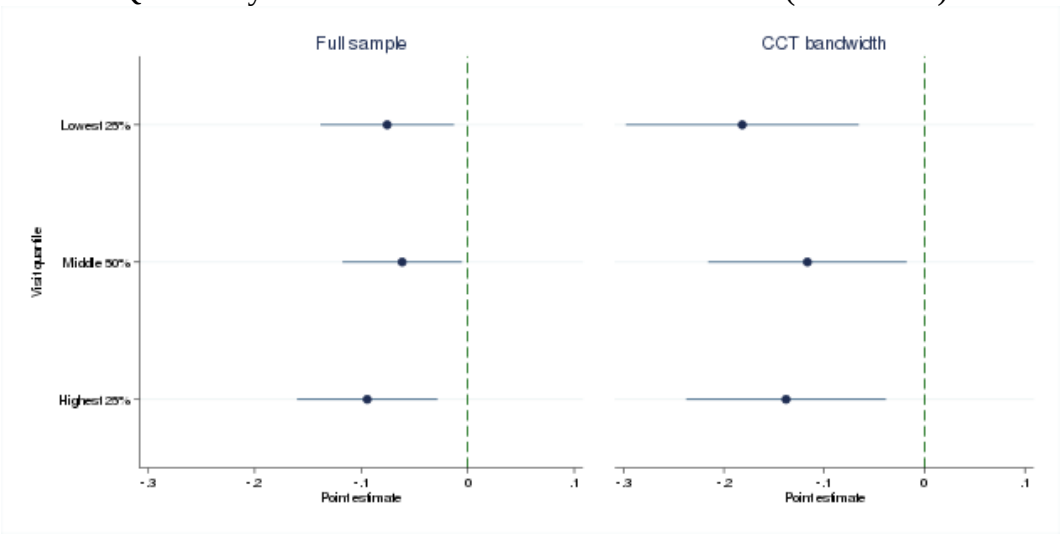
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A-1. Estimated effects by coaching dosage

Panel A: Quartile by instructional coach visit count (cumulative)

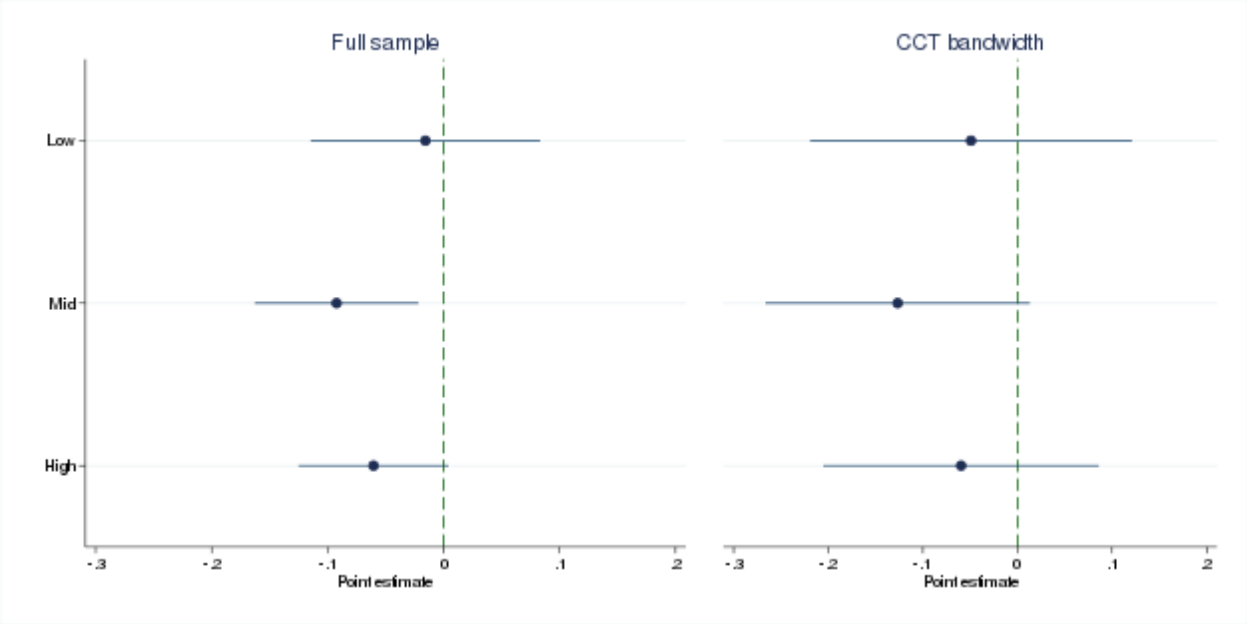


Panel B: Quartile by school transformation coach visit count (cumulative)



NOTE: 2SLS estimates from fuzzy RD using triangular kernel with separate treatments for schools in the bottom quartile of number of visits, middle 50% of number of visits, and top quartile of number visits. Quartiles by school level. All first-stage test statistics are greater than What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage. Preferred CCT bandwidths from fuzzy test score models. Standard errors clustered at the school level.

Figure A-2. Estimated effects by fidelity of implementation quartile



NOTE: 2SLS estimates from fuzzy RD using triangular kernel with separate treatments for three different categories of FOI. Low group has mean score of less than 2 on 1-4 scale, mid group has mean score of 2 to less than 3, and high group has mean score of 3 or above. Sample restricted to schools with full FOI index scores (14 of 78 ITT schools are missing index scores). All models include lagged score and subject on the right side, with math as the reference category. All first-stage test statistics are greater than What Works Clearinghouse (2017) recommended minimum size of 4 for a sufficiently strong first stage. Preferred CCT bandwidths from fuzzy test score models. Standard errors clustered at the school level.