# The Next Generation of State Reforms to Improve their Lowest Performing Schools:

# An Evaluation of North Carolina's School Transformation Intervention

Gary T. Henry [1]

Erica Harbatkin [2]

[1] Dean, College of Education & Human Development, University of Delaware, gthenry@udel.edu

[2] Postdoctoral Research Associate, Michigan State University, erica.s.harbatkin@vanderbilt.edu

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 achievement 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. While the increased teacher turnover in NCT schools in 2017 opens the possibility that reform schools were intentionally replacing less effective teachers with more effective ones, our analysis does not support that strategic staffing occurred. The negative effects on student achievement appear related to variable timing of implementation of one of the few required components for serving low-performing schools under ESSA—a comprehensive needs assessment which leads to comprehensive school improvement plans. These findings may serve as a cautionary tale for states planning low-performing school reforms under ESSA.

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.

# 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 et al., 2017; Zimmer et al., 2017), although some studies have found negative or null effects (Dickey-Griffith, 2013; Dragoset et al., 2017; Heissel & Ladd, 2018; Henry et al., 2015). In many of the turnaround efforts that have been shown to be effective, strategic staffing—which involves replacing less effective teachers by recruiting, hiring, developing and retaining more effective teachers—seems to have played a role in successful turnaround, as we discuss later. However, reforms that successfully recruit and retain effective teachers from other schools may also induce general equilibrium effects, which lowers performance in the schools from which the teachers transferred when the supply of effective teachers is limited (Kho et al., 2019).

Under ESSA, the federally mandated turnaround models have faded into the past along with additional dedicated funding for turnaround. The school reform interventions implemented under No Child Left Behind (NCLB) waivers yielded less consistent effects on student achievement, with just one study finding positive effects, three with null effects, and one with negative effects (Atchison, 2020; Bonilla & Dee, 2017; Dee & Dizon-Ross, 2019; Dougherty & Weiner, 2017; Hemelt & Jacob, 2017, 2018). Within the same intervention, heterogeneous

effects have been driven by variation in implementation of reforms (Dougherty & Weiner, 2017; Strunk et al., 2016).

Turnaround under ESSA will share more in common with NCLB waivers and similar district- and state-initiated reforms than RttT and SIG for two reasons. First, states will have flexibility in how they reform 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 four 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 engage in strategic replacement of teachers by hiring more effective replacement teachers and losing less effective teachers?
4. Is variation in implementation associated with differences in outcomes?

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. Moreover, exiting teachers were not less effective than those exiting control schools, and entering teachers were not more effective than those entering control schools, undermining the strategic replacement hypotheses. The negative effects appear to be associated with the timing of the comprehensive needs assessment, a required component of reform under ESSA,

which was intended to precede the state's supports to its lowest performing schools. These findings may serve as a cautionary tale for how states engage in mandated comprehensive school improvement (CSI) under ESSA.

## School Turnaround

Prior research has shown substantial heterogeneity in the effects of whole school reform efforts (Gross et al., 2009). SIG and RttT introduced school turnaround to the federal school reform agenda in 2008. Turnaround was distinguished by the urgency to 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—many of which included substantial staff replacement and practices aimed at recruiting, retaining and developing effective teachers—produced strong positive effects on student achievement in Massachusetts, Tennessee (local Innovation Zones), Ohio, and California (Carlson & Lavertu, 2018; Dee, 2012; Henry et al., 2020; Papay & Hannon, 2018; Schueler et al., 2016; Strunk et al., 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 heterogeneity within interventions. Turnaround in North Carolina, Tennessee, 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).

Heterogeneity of the effects of school reform models continued under NCLB waivers, with fewer positive effects than RttT and SIG. Of four states with evaluations of waiver reforms, one—Kentucky—produced positive effects on student achievement, which the authors attributed to the state's focus on reducing achievement gaps combined with a clearly articulated set of reform activities from the state (Bonilla & Dee, 2017). Reforms in New York, Michigan, Rhode Island, and Louisiana produced either null or negative effects on student achievement (Atchison, 2020; Dee & Dizon-Ross, 2019; Dougherty & Weiner, 2017; Hemelt & Jacob, 2017, 2018).

The mixed effects of interventions under both RttT/SIG, NCLB waivers, and locally initiated reforms 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, New York, 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 measures, and empirical strategy, followed by the findings, and a series of validity checks. 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 most recent 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 et al., 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.[1] An early adopter of turnaround because of a 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. 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 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 "unpacking" the findings by state facilitators who discussed CNA findings with school staff. The 1.5-day unpacking process consisted of three elements: (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. The unpackings generally occurred during the summer following the school year of the CNA, although there was variation in when and whether schools received them.

Table 1 ABOUT HERE

The CNA and unpacking were intended as the springboard from which school improvement planning and subsequent turnaround activities would occur. Early in the school year, 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. To that end, the timing of the CNA was central to each school's turnaround plans. Because CNA findings were intended to inform the SIP, the NCT model relied on the CNA occurring prior to the school improvement planning process. However, planning delays combined with limited state resources precluded the state from carrying out CNAs in a manner consistent with the theory of action, as Table 1 outlines. Only about one-fifth of treatment schools received CNAs before developing their SIPs, though these occurred as many as three

years prior to the intervention when schools could have had different staff and different needs. Meanwhile, one-third of treatment schools received CNAs during the spring semester of the first year of the intervention after they had already developed their SIPs.

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 improvement intervention was the coaching that followed, with the goal of building leadership capacity through school transformation coaching and teaching capacity through instructional coaching. NCT was intended as a tailored intervention in which coaches were responsive to school, principal, and teacher needs. Not all schools received both school transformation and instructional coaching, and there was wide variation in the number, content, and structure of the visits. While the average treatment school received 45 instructional coach visits and 25 school transformation coach visits over the three-semester period, Table 2 highlights the considerable variation across schools. Of particular relevance is the wide range of dosage; treatment schools receiving very few coaching visits experienced fundamentally different exposure to services than treatment schools receiving weekly visits.

Table 2 ABOUT HERE

The intervention did not closely mirror any of the four previous federal school turnaround models. 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 focused resources on developing existing staff rather

than on recruiting and retaining effective staff. 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 "top-down" turnaround models. This evaluation can therefore help to inform state turnaround policy under ESSA, under which states are required to undertake needs assessments and school improvement plans in their lowest performing schools and have the flexibility to implement school turnaround interventions that look like NCT.[2]

## Sample

The sample includes all North Carolina schools that the state 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 standards, 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 assigned to 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. In order to mitigate bias that would arise from these always-takers and never-takers, our inferences apply only to compliers. 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 and Measures

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 scores. 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 2015-16 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 mobility was intentional and strategic, we draw from two lagged measures of teacher effectiveness. Subject-specific 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.[3]

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.[4]

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).

**Implementation**

We examine three implementation measures to determine whether variation in implementation was associated with differences in outcomes—focusing specifically on dimensions of implementation with substantial variation across schools. These three dimensions

are the timing of the CNA, the presence of a CNA unpacking, and the dosage of coaching. We collected measures for each of these variables directly from the state and validated them through site visits to treatment schools and phone interviews with control school principals. The state did not provide coaching to control schools, and, according to documents provided by the state, did not undertake comprehensive needs assessments in control schools.

**CNA timing.** We drew from the state's CNA and unpacking calendar to categorize schools by CNA timing. The state prioritized scheduling CNAs in schools that had not received one in the three years prior to the beginning of supports. We therefore placed schools in four categories according to CNA timing: (1) CNA was more than three years before services began or did not occur at all—these are schools that did not receive the CNA component of the intervention on the timeline set by the theory of change; (2) CNA was during the three years before the start of services—these are schools for which the CNA was implemented as specified by the theory of change timeline but for which CNA findings may have no longer been relevant because of staff turnover or changing school dynamics; (3) CNA was in spring 2016—these are schools that received a CNA during the intervention but during the school year and after they would have already developed and begun to implement their school improvement plans; and (4) CNA was during the 2016-17 school year, the second year of supports. 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 or potentially lead to a disruption of school improvement efforts already in progress under the SIP developed at the beginning of the school year.

**CNA unpacking.** Also drawing from the state CNA and unpacking schedule, we created a single dichotomous measure denoting whether or not the school received an unpacking.[5] Forty-

nine schools received an unpacking and 26 did not. Because the unpacking was intended to build from the CNA and provide schools with a path forward for the school improvement plan, we hypothesized that schools that received unpackings may have been able to develop more targeted improvement plans leading to better outcomes.

**Coaching dosage.** Using coaching reports provided by the state, we counted the number of school transformation, instructional, and total coaching visits carried out during the intervention period. The dosage measure takes three values: high dosage schools (top quartile), mid-dosage schools (middle 50%), and low dosage schools (bottom quartile). 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 rather than building the capacity of principals and teachers.

### Covariates

School-level variables include minority percentage, economically disadvantaged percentage, per pupil expenditures (PPE) and PPE squared, enrollment (average daily membership, or 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

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

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 $\varepsilon$ is an idiosyncratic error term clustered at the school level. In a second set of models, we add vectors of school-level covariates, ***S'***, and individual-level covariates, ***K'***, to increase precision. The individual-level covariates are student level in models predicting student test score 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. $\beta_1$ 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.[6]

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

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

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

$$y_{is} = \beta_0 + \beta_1 (\widehat{NCT})_s + \beta_2 f(GLP)_s + \beta_3 I(GLP < 0)_s \times f(GLP)_s + \pi \boldsymbol{S}'_s + \rho \boldsymbol{K}'_i + \varepsilon_{is} , \quad (3)$$

where the predicted outcome, *y*, for student or teacher *i*, is a function of the predicted *NCT* indicator, and the model 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 $\beta_1$ 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 et al., 1996). In other words, having a performance composite below the threshold needs to affect student and teacher outcomes only through its effect on the likelihood of receiving turnaround services. The ITT estimates would not be subject to the same bias. While we cannot strictly test

whether the exclusion restriction is met, we did survey principals in treatment and control schools to examine implementation in treatment schools and whether control schools may have received similar services. We draw from these data to descriptively examine whether (a) non-takers of assignment to treatment (no-shows) report similar levels of coaching to comparison group compliers, and (b) always-takers report similar levels of coaching to treatment group compliers. Using this descriptive analysis, we do not find differences between these groups in the probability of reporting receipt of school transformation or instructional coaching (Table A-1 provides crosstabulations of these groups with chi-square tests). While this approach to examining the exclusion restriction is limited to the schools for which we have survey responses on the relevant questions, we believe that these findings, combined with the similarity of the ITT and TOT estimates, suggest that any violation of the exclusion restriction that we are unable to observe would likely be minor and have only a negligible effect on the TOT estimates.

We stack all subjects in our main student achievement specification but also include separate models for math, reading, and science in the appendix. The lagged test score on the right-hand side of the equation is from one year prior for fourth- through eighth-grade math and reading. For high schools, the lag is 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.[7] 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 at the cutoff.

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,[8] 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.[9] 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.[10]

**Teacher effectiveness**

After estimating the effects of the intervention on student achievement 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 whether the effects of NCT on teacher turnover and new-to-school teachers differ between more and less effective teachers. We define three treatment groups—high, mid, and low effectiveness teachers in NCT schools—using EVAAS scores, NCEES standard 1, and NCEES standard 4. To estimate the effects of NCT on each of these three groups of teachers, we implement analyses following the fuzzy RD framework with equations that are analogous to equations (2) and (3) predicting two dichotomous outcomes—turnover and being new to school—using three separate treatment groups rather than a single treatment. Specifically, to test for heterogeneity between teachers in each of three levels of effectiveness, we interact assignment to treatment with each of the three categories of teacher effectiveness. In order to estimate within-group differences between NCT and control schools, we also include indicators for high and low-effectiveness teachers. Because we have three treatment groups, we estimate three first-stage models predicting turnaround status within each of the three

groups based on teacher effectiveness category. Equation 4 represents the equation for the highly effective group of teachers. In the other two first-stage equations we substitute *HighlyEffective* with *MidEffectiveness* and *LowEffectiveness* indicators. Otherwise, all three first stage equations are identically specified.

$$\begin{aligned}\Pr\,(NCT_s|HighlyEffective_i) &= \alpha_0 + \alpha_1(HighlyEffective_i \times I(GLP<0)_s) + \alpha_2 f(GLP)_s \\ &+ \alpha_3 I(GLP<0)_s \times f(GLP)_s + \alpha_4 HighlyEffective + \varepsilon_{is}\end{aligned} \quad (4)$$

The first-stage outcome in equation 4 is the predicted probability of being in a treated school for highly effective teachers. In other words, the first stage estimates the probability of being in a treated school, conditional on the school's assignment to treatment and the teacher being in the highly effective category. The coefficient estimate represented by $\alpha_1$ provides the estimated effect of a teacher being in a particular effectiveness group in a school below the cutoff on the probability of treatment. The first-stage equations produce three separate predicted variables to carry into the second stage—one for each of the three treatment groups represented by teachers of high, mid, or low effectiveness. We then include the fitted values of the dependent variables from the three first-stage equations as predictors in the second stage:

$$\begin{aligned}y_{is} = \beta_0 &+ \beta_1(NCT_s \mid HighlyEffective_i) + \beta_2(NCT_s \mid MidEffectiveness_i) \\ &+ \beta_3(NCT_s \mid LowEffectiveness_i) + \beta_4 f(GLP)_s \\ &+ \beta_5 I(GLP<0)_s \times f(GLP)_s + \beta_6 HighlyEffective_i \\ &+ \beta_7 LowEffectiveness_i + \varepsilon_{is}\end{aligned} \quad (5)$$

The outcome (turnover or new to school, represented as *y*) for teacher *i* in school *s* is estimated using the same approach as equation 3, but in equation 5, the three teacher effectiveness predicted values allow separate within-effectiveness-group estimates of the probability of turnover or being new to school. The $NCT_s \mid HighlyEffective_i$ variable is the predicted value of the dependent variable from Equation 4 above, so $\beta_1$ represents the complier-adjusted local

average treatment effect for highly effective teachers in NCT schools. The $NCT_s \mid MidEffectiveness_i$ and $NCT_s \mid LowEffectiveness_i$ variables represent the predicted values of the dependent variables from the two parallel first-stage equations. We present estimates from these models without additional covariates, though the estimates are robust to inclusion of school- and teacher-level covariates.

Evidence of strategic staffing would be apparent when examining the $\beta_1$ and $\beta_2$ coefficient estimates. In the model estimating effects on turnover, a negative and significant estimate on the highly effective group ($\beta_1$) would provide evidence that treatment schools retained more effective teachers than control schools, while a positive and significant estimate on the low effectiveness $\beta_3$ would provide evidence that more of the less effective teachers left treatment schools than the control group schools. In the model predicting new-to-school teachers, a positive and significant estimate on the highly effective group ($\beta_1$) would provide evidence that treatment schools hired more effective teachers, while a negative and significant estimate on low effectiveness group ($\beta_3$) would provide evidence that treatment schools hired fewer ineffective teachers relative to control schools.

**Implementation**

We use a similar approach to test for heterogeneous effects by each of the three dimensions of implementation, replacing the teacher effectiveness group with the appropriate implementation category (by CNA timing group; whether or not the school received an unpacking; and high, mid, and low coaching dosage) and the outcome with student achievement. While equations (4) and (5) represent a traditional moderation approach comparing groups of teachers in treatment schools with similarly effective teachers in control schools, the implementation analysis simply compares the performance of students in groups of treatment

schools with the performance of students in all control schools within the bandwidth. We take this approach because control schools did not have CNAs scheduled during the study period and therefore could not be placed into subgroups based on CNA timing. We illustrate the empirical approach with one of the three dimensions of implementation that we examined, the timing of the CNA. The first-stage model for the group of schools that did not receive CNAs or received one prior to 2014 therefore takes the form

$$\begin{aligned}\Pr(NCT|NoCNA)_s &= \alpha_0 + \alpha_1\big(NoCNA \times I(GLP < 0)\big)_s \\ &+ \alpha_2\big(2014or2015CNA \times I(GLP < 0)\big)_s \\ &+ \alpha_3\big(Spring2016 \times I(GLP < 0)\big)_s + \alpha_4\big(201617 \times I(GLP < 0)\big)_s \\ &+ \alpha_5 f(GLP)_s + \alpha_6 I(GLP < 0)_s \times f(GLP)_s + \alpha_7 Score_{it-1} + \varepsilon_{is}\end{aligned} \quad (6)$$

where the first-stage outcome is the predicted probability of being in a treated school that did not receive a CNA or received one prior to 2014. Here, we estimate separate first-stage equations for each implementation group; in other words, for the CNA timing analysis, we estimate three additional first-stage equations predicting the probability of being in a treated school that received a CNA in 2014 or 2015, being in a treated school that received a CNA in spring 2016, and being in a treated school that received a CNA in the 2016-17 school year. We carry the predicted values from each of the four first-stage equations into the second stage:

$$\begin{aligned}Score_{is} = \beta_0 &+ \beta_1(NCT|NoCNA)_s + \beta_2(NCT|2014or2015CNA)_s \\ &+ \beta_3(NCT|Spring2016CNA)_s + \beta_4(NCT|2016-17CNA)_s \\ &+ \beta_5 f(GLP)_s + \beta_6 I(GLP < 0)_s \times f(GLP)_s + \beta_3 Score_{it-1} + \varepsilon_{is}\end{aligned} \quad (7)$$

where the test score for student *i* in school *s* is a function of each of the predicted probabilities of being in the four categories of treatment by CNA timing from the first-stage equations, the forcing variable, an interaction between the forcing variable and being assigned to treatment, the student's lagged test score, and vectors of school and student covariates. The coefficient

estimates on the four separate treatments represent the estimated effect of being in an NCT school that received a CNA in a particular time period on student achievement, relative to students in all control schools at the cutoff. Estimates on $\beta_1$ through $\beta_4$ that are different from one another would provide evidence that variation in implementation was associated with differences in outcomes.

We consider the implementation analyses correlational rather than causal because schools were not randomly assigned to implementation variation such as CNA receipt in a particular time period. We do examine whether the four subgroups of schools that vary with CNA timing appear to have different effects at the time of assignment to treatment. To do so, we regress the school baseline performance composite on implementation groupings. We do not find that the CNA timing groups are significant predictors of baseline performance. These results, along with results for other implementation groupings, are provided in Table A-2.[11]

## Results

We find consistent evidence that NCT had a negative effect on student achievement in 2017 and neither a positive or negative effect in 2016. 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 has a negative effect, -0.12 standard deviations, on test scores 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 -0.13 in 2017 in our preferred model. The similarity in terms of both magnitude and significance of the ITT and TOT estimates suggest that the noncompliance has a negligible effect on the ITT estimates. We proceed by showing TOT estimates from the fuzzy RDs in the remainder of the manuscript.

Similar to the ITT estimates, the 2017 TOT estimates displayed in Table 5 are consistently negative and significant across the three bandwidths and with and without covariates. In the analytical sample defined by the narrowest bandwidth, we find a negative effect of NCT in 2016. This pattern of effects is similar when we estimate within alternative bandwidths and using the full sample. The effects in 2017 are consistently negative and significant and the effects in 2016 are only significant when the analytical sample is defined by

narrower bandwidths calculated using the bandwidth selection procedure described in Imbens & Kalyanaraman (2009). [12]

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 *F*-statistic on the treatment indicator to show that the instrument is a sufficiently strong predictor of treatment is 16 (What Works Clearinghouse, 2017). All first-stage *F*-statistics exceed this criterion as shown in Table 5.

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 student achievement (Table A-4). We also find qualitatively similar results when we estimate on test score levels rather than conditioning on lagged achievement values, shown in Table A-5, 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 TOT estimates in elementary schools (Table A-6).

Figure 4 ABOUT HERE

**Teacher turnover.** We also find evidence that teachers in NCT schools were more likely to turn over in 2017 but neither more nor less likely to turn over in 2016, displayed visually in Figure 4 and numerically in Table 6. In 2017, teachers in NCT schools were 22.5 percentage points more likely to turn over than control school teachers. These estimates are consistent across bandwidths and robust to the inclusion of covariates (Table A-7).[13]

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, 2020; Ronfeldt, Loeb, & Wyckoff, 2013), strategically replacing lower performing teachers with more effective teachers can have positive effects—especially in very low performing schools (Adnot et al., 2017; Henry et al., 2020; 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 that negative compositional effects were likely to have produced the negative effects on student achievement in 2017. If these negative effects were driven by turnover of highly effective teachers paired with replacement by less effective 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 compositional effect of turnover drove negative effects on student achievement.

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 NCT 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 in the following year, 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.

**Implementation.** While we do not find evidence of dosage effects or the availability of an unpacking after the CNA (Figure A-2 and Figure A-3), we do find suggestive evidence that the negative effects on student achievement in 2017 appear to be concentrated in three categories of CNA timing. In particular, Figure 5 shows that the negative effects in 2017 occurred 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 when coaching services were being delivered. We observe null effects in the

17 schools that received CNAs in the 2016-17 school year. While we believe the pattern of effects is informative, this finding should nevertheless be interpreted with caution given the overlapping confidence intervals for the four subgroup effects.\

Figure 5 ABOUT HERE

Qualitative data collected as a part of the overall evaluation provides some context for interpreting these results. Descriptively, the largest negative effects appear in schools that did not receive a CNA within a period useful for school improvement planning (more than two years before the NCT services began, if ever). The intervention delivered in these schools effectively undermined the theory of change, which predicated the reform strategy on an in-depth assessment of school needs 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 were surprised and upset 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 the summer of 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. During interviews, school staff reported that they viewed CNAs conducted during the 2016-17 school year more favorably and our findings show no negative effects among this group of schools.

## 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. Additionally, graphical analysis shows no evidence of manipulation of the forcing variable around the cutoff.[14] 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.[15]

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-3 and Table A-7).

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 F-statistics on the treatment indicator are above the minimum recommended threshold of 16 (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.

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 sample is within the recommended .25 standard deviation units of the control sample on key covariates associated with school performance, conditional on the forcing variable, within the 2016 and 2017 preferred CCT bandwidths. Table 8 shows effect sizes from a series of models estimating the baseline (2015) covariate value using the forcing variable and a triangular kernel within the preferred bandwidth for each year. None of the treatment effect size estimates exceeds .25 standard deviation units, which demonstrates the treatment and control samples are balanced on observed covariates within the 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-9).

Another assumption of the RD design is that student selection into or out of the treatment schools in response to the intervention is minimal. To test this assumption, we examine whether the demographics of NCT schools changed in response to the intervention in 2016 or 2017. Specifically, we estimate an RD on a set of school-level demographic characteristics in each of the two years of treatment within the optimal bandwidth. In the presence of student selection in and out of treatment schools, we would observe significant effects of NCT on these school-level demographics variables. Table A-10 shows there is no evidence for these selection effects.

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

## Discussion

We find that NCT reduced student achievement and increased teacher turnover and that the negative effects on student achievement may be associated with the timing of the CNA. 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, we find no evidence that strategic staffing occurred. Treatment schools experienced higher turnover across all levels of teacher effectiveness, with ineffective teachers no more likely to turn over than more effective

teachers in their schools. Given NCT's largely rural context, this finding underscores the challenges associated with recruiting and retaining effective teachers in rural schools, which are unlikely to have robust educator labor markets from which to draw. Turnaround efforts that rely on strategic staffing may be less effective in rural contexts if they fail to counteract these labor market challenges with financial incentives that were a part of some effective turnaround efforts or other effective approaches aimed at recruiting and retaining effective teachers. Our findings that NCT schools did not recruit more effective teachers provide evidence against the possibility of a general equilibrium effect of targeted turnaround on nearby schools such as what occurred in Tennessee's iZones (Kho et al., 2019). However, unlike the iZones, which comprised urban schools, those dynamics may be expected to play out differently in rural schools that need to recruit teachers from outside the local area.

Under NCT, 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 focusing on schoolwide processes and practices such as establishing a supportive and collaborative environment, NCT prioritized coaching to develop individual teacher skills and capacity in schools where, on average, the entire staff turns over every three years. 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. Strategic staffing is less likely to be an effective 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, especially without regulations and funds to support incentives for effective teachers to transfer into and remain in these low-

performing schools. In fact, teachers in NCT schools expressed being stigmatized from working in a school labeled as “low performing.”

While we cannot know for certain whether the first two years of NCT laid the groundwork for incremental 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 the 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.

It is possible that targeting all schools in the bottom 5 percent produced negative effects by spreading resources too thin. Specifically, providing limited, inconsistent supports 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 increasing 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 needs assessment timing and finding the resources, both human and financial, to conduct the needs assessments prior to the school improvement planning and implementation. Such efforts might require a planning year and additional human resources prior to initiating comprehensive services.

Three limitations are relevant to interpreting these findings. First, the regression discontinuity design focuses intentionally on schools around the eligibility cutoff in order to minimize threats to internal validity related to baseline differences between schools. While the findings are consistent across a wider set of bandwidths, the RD estimates represent the estimated effects of NCT for a narrow band of schools around the cutoff and the generalizability of the estimates is limited by the focus on these schools. Second, 21 of the schools receiving NCT services also received turnaround services under the state's RttT grant. Because these schools were in the bottom 5 percent in two different rounds of identification, it is possible that they may be more resistant to turnaround efforts and that the negative effects stem in part from that resistance. Finally, the generalizability of these findings should be considered in the context of the sample. The implementation of 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, 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 promote strategic staffing practices. These strategies were included as part successful turnaround models in other states.

## Conclusion

As states implement plans to support their lowest performing schools under ESSA, our findings suggest that school reform without intentional disruption of the status quo or supplemental resources has the potential to hinder student achievement and increase

unintentional teacher turnover. This analysis also suggests that direct service provision without the backing of an influx of funding may not be a viable turnaround strategy across the entire set of schools in the bottom 5 percent in each state.

While these findings provide some descriptive evidence to explain the mechanisms underlying the negative effects of NCT on student achievement, future research could examine factors that may mediate or suppress the effects of interventions to improve the lowest performing schools. Such factors may include implementation fidelity and quality, school morale, and school climate.

[1] 21 of the 75 NCT schools had received services under TALAS.

[2] 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.

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

[4] 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.

[5] We did not categorize schools by unpacking timing as we did by CNA timing because unpacking timing overlapped closely with CNA timing. Findings using unpacking timing are similar to those using CNA timing.

[6] We use the rdrobust package in Stata to estimate the optimal bandwidths and the RD models (Calonico et al., 2017).

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

[8] 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.

[9] 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.

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

[11] While we do not find significant effects on our primary implementation analysis, which is focused on CNA timing, we do find significant effects in two implementation groupings that we show in the appendix. First, having an unpacking in 2014 or 2015 was associated with a lower baseline performance composite than having no unpacking. Second, schools in the highest quartile of instructional coaching visit dosage had lower predicted baseline performance composites than schools in the middle 50% of instructional coaching visit dosage.

[12] 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.

[13] Results from the full analytical sample and alternative IK bandwidths are presented in in Tables A-7 and A-8. While a weak first stage in the 50 percent IK bandwidth for 2017 precludes valid inferences for the TOT estimate

within this bandwidth, a sharp specification finds significant increases in teacher turnover in the narrowest bandwidth and across other bandwidths (Table A-8).

[14] We show the density of the forcing variable across the full sample of eligible schools in Figure A-1

[15] 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

Atchison, D. (2020). The Impact of Priority School Designation Under ESEA Flexibility in New York State. *Journal of Research on Educational Effectiveness*, *13*(1), 121–146. https://doi.org/10.1080/19345747.2019.1679930

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

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *Stata Journal*, *17*(2), 372–404.

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). National Bureau of Economic Research. 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., Boyle, A., Upton, R., Tanenbaum, C., & Giffin, J. (2017). *School Improvement Grants: Implementation and Effectiveness. NCEE 2017-4013*. National Center for Education Evaluation and Regional Assistance. 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). National Bureau of Economic Research. 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*. http://cerenc.org/wp-content/uploads/2015/09/ES-FINAL-Final-DST-Report-9-3-15.pdf

Henry, G. T., Pham, L. D., Kho, A., & Zimmer, R. (2020). Peeking Into the Black Box of School Turnaround: A Formal Test of Mediators and Suppressors. *Educational Evaluation and Policy Analysis*, *42*(2), 232–256. https://doi.org/10.3102/0162373720908600

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* (NCEE 2008-4020; IES Practice Guide). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/wwc/PracticeGuide.aspx?sid=7

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

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

Kho, A., Henry, G. T., Zimmer, R., & Pham, L. (2019). *General Equilibrium Effects of Recruiting High-Performing Teachers for School Turnaround: Evidence from Tennessee* (EdWorkingPaper: 19-64). Annenberg Institute at Brown University. https://edworkingpapers.com/ai19-64

Papay, J., & Hannon, M. (2018, November 8). *The Effects of School Turnaround Strategies in Massachusetts*. 2018 APPAM Fall Research Conference: *Evidence for Action: Encouraging Innovation and Improvement*, Washington, D.C. 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). National Bureau of Economic Research. http://www.nber.org/papers/w21895

Stock, J. H., & Yogo, M. (2002). *Testing for Weak Instruments in Linear IV Regression* (Working Paper No. 284). National Bureau of Economic Research. 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

What Works Clearinghouse. (2017). *Standards Handbook. Version 4.0.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. 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<br>16 math, 18 ELA, 12 science, 33 non-subject-specific | 56 total |
| Number of visits | Range: 0-137<br>Mean: 45.36 | Range: 0-63<br>Mean: 25.28 |
| Visits per teacher | Range: 0-15.75<br>Mean: 1.83 | Range: 0-3.82<br>Mean: 1.03 |

Source: DPI coaching reports for three semesters from spring 2016, fall 2017, and spring 2017.

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<br>(0.11) | 0.1<br>(0.23) |
| Suburb | 0.0<br>(0.00) | 0.1<br>(0.22) |
| Town | 0.1<br>(0.25) | 0.1<br>(0.30) |
| Rural | 0.9<br>(0.27) | 0.8<br>(0.41) |
| *School level* | | |
| Elementary | 48.7<br>(50.31) | 57.3<br>(49.56) |
| Middle | 35.9<br>(48.28) | 33.6<br>(47.33) |
| High | 15.4<br>(36.31) | 9.1<br>(28.80) |
| *Student achievement* | | |
| 2015 performance composite (centered) | -5.1<br>(4.31) | 9.8<br>(5.43) |
| EVAAS growth score | 68.6<br>(10.36) | 68.5<br>(10.64) |
| *Teacher qualifications* | | |
| Percent novice teachers | 32.5<br>(12.57) | 26.8<br>(12.27) |
| Percent National Board Certification | 7.7<br>(4.64) | 11.5<br>(7.24) |
| *Student demographics* | | |
| Minority percent | 84.7<br>(12.44) | 60.2<br>(21.70) |
| Economically disadvantaged | 82.2<br>(12.12) | 72.7<br>(14.66) |
| *School characteristics* | | |
| Per pupil spending | 10217.7<br>(2264.00) | 9426.0<br>(1863.50) |
| Average Daily Membership | 429.2<br>(172.67) | 498.8<br>(226.64) |

NOTE: Means and standard deviations on baseline measures based on 331 eligible schools.

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

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 scores)***

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 *F*-stat | 120.78 | 113.64 | 35.28 | 36.84 | 213.16 | 213.16 |
| 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 *F*-stat | 81.72 | 88.74 | 29.81 | 49.14 | 222.01 | 211.41 |
| 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 *F*-stat | 31.47 | 32.60 | 7.84 | 6.45 | 74.13 | 76.74 |
| 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 *F*-stat | 24.01 | 24.80 | 6.86 | 5.52 | 66.75 | 69.06 |
| 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 F statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 16 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 F-statistics are greater than the What Works Clearinghouse (2017) recommended minimum size of 16 for a sufficiently strong first stage, except for the test statistic for TOT x average EVAAS in Model 6, which is 10.24.

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

**Table 8. Sample balance on standardized variables, conditional on forcing variable within optimal bandwidths**

| | 2016 | | 2017 | |
|---|---|---|---|---|
| | β | SE | β | SE |
| Female | 0.034 | 0.033 | 0.030 | 0.029 |
| White | -0.192 | 0.014 | -0.145 | 0.013 |
| Black | 0.224 | 0.030 | 0.191 | 0.026 |
| Hispanic | -0.059 | 0.030 | -0.131 | 0.027 |
| Other race | 0.106 | 0.008 | 0.103 | 0.007 |
| Disabled | 0.057 | 0.032 | 0.046 | 0.029 |
| Gifted | 0.184 | 0.027 | 0.127 | 0.025 |
| Limited English proficiency | -0.101 | 0.033 | -0.138 | 0.030 |
| Over-age for grade | 0.248 | 0.032 | 0.208 | 0.029 |
| Nonstructural transfer in | 0.021 | 0.034 | 0.030 | 0.030 |
| Economically disadvantaged | -0.088 | 0.032 | -0.077 | 0.028 |

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

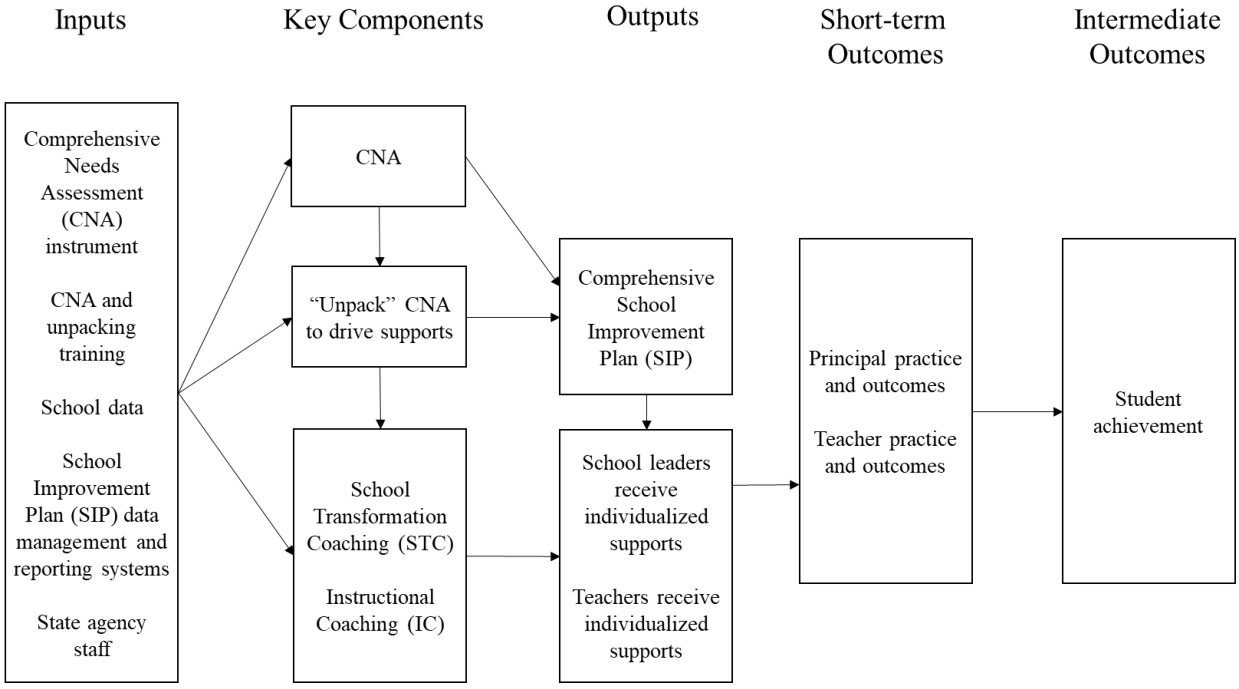
**Table 9. Attrition**

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

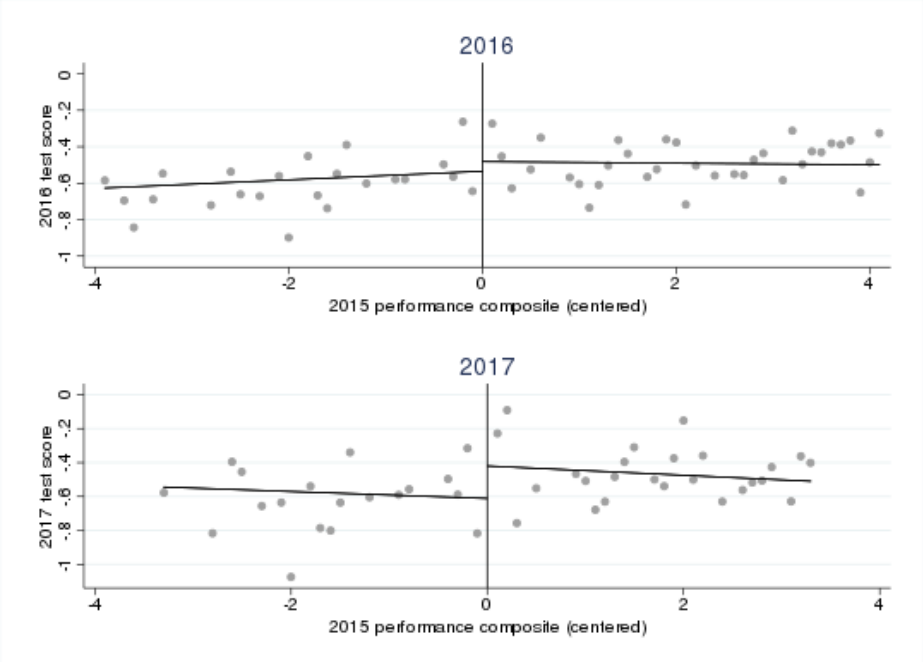
NOTE: Estimates from linear probability model predicting attrition at the school level and controlling for the forcing variable within the optimal CCT bandwidths and with a triangular kernel.

# Figures

## Figure 1. North Carolina Transformation Logic Model

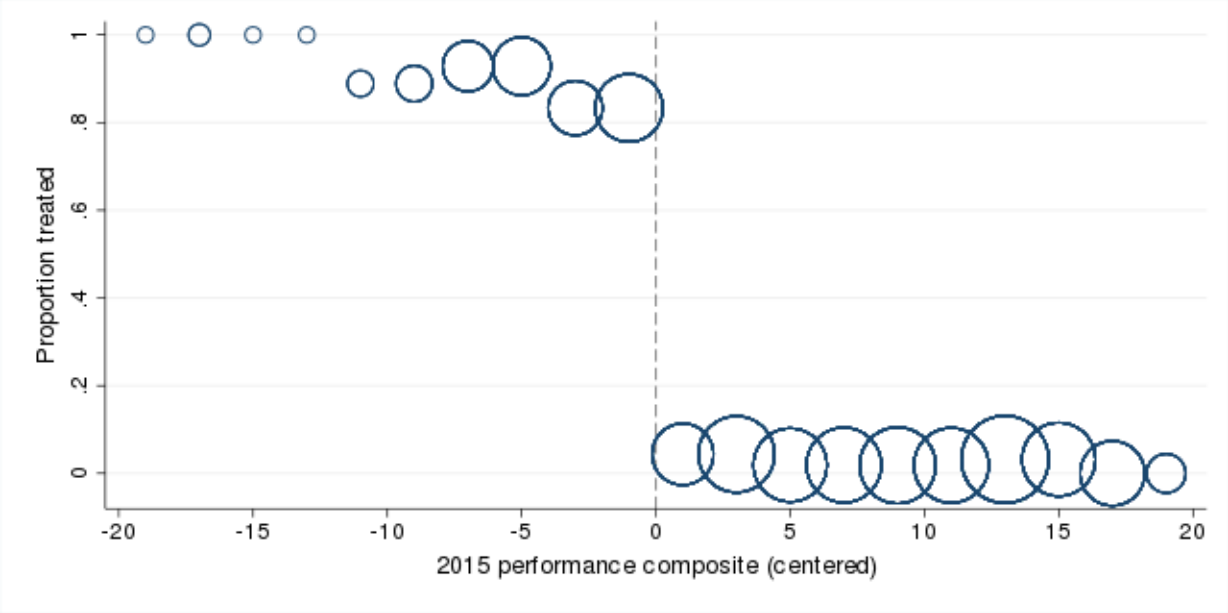



## Figure 2. Student achievement by distance from assignment threshold



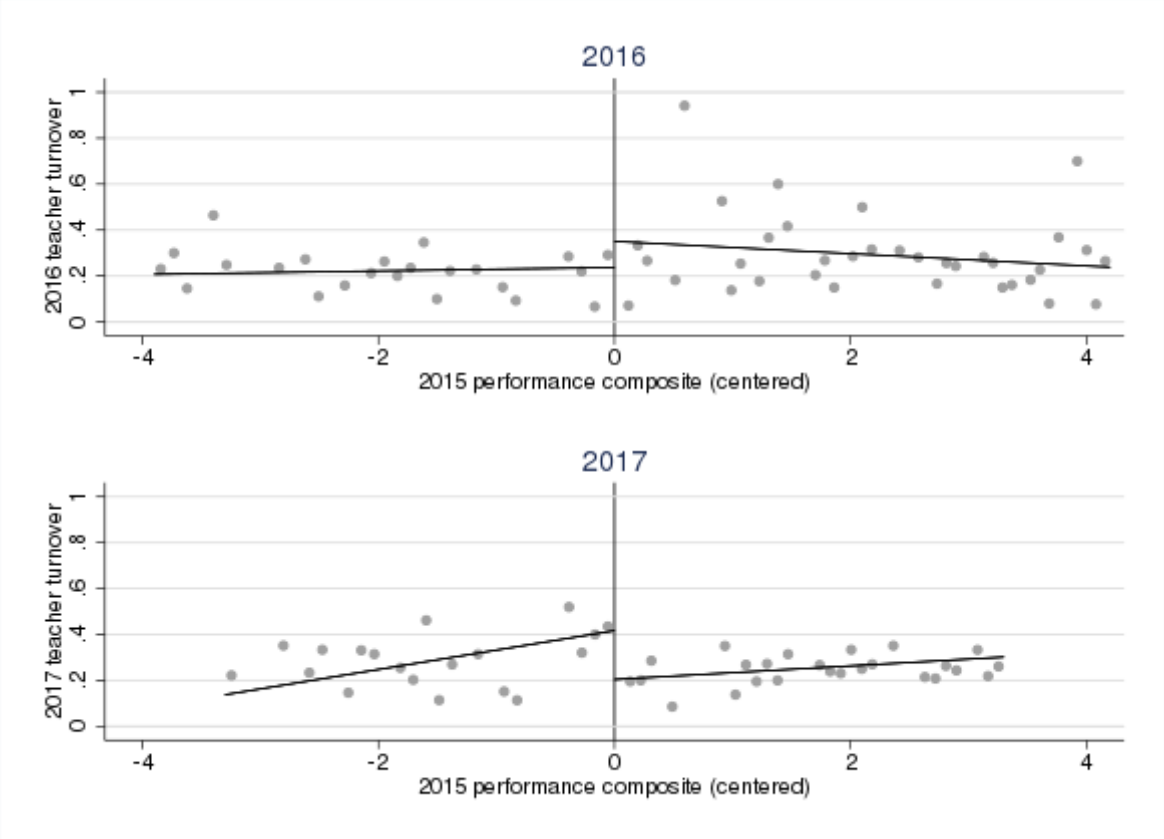

NOTE: Markers represent bin averages within CCT bandwidths and lines are linear fit. Estimation using triangular kernel within preferred CCT bandwidth, with average bin width of .006 to left of cutoff and .007 to right of cutoff in 2016, and .007 to left of cutoff and .010 to right of cutoff in 2017.

**Figure 3. Proportion treated by forcing variable**



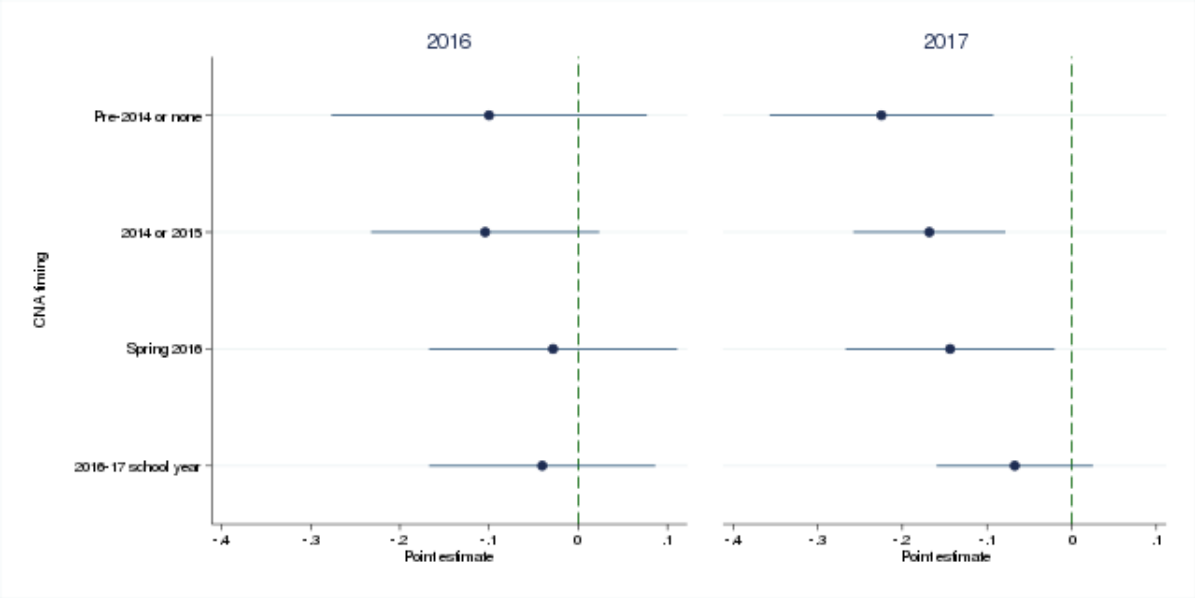

NOTE: Markers represent bin averages. Bin width is 2. Marker sizes weighted by number of schools in bin.

**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. 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 *F*-statistics are greater than the What Works Clearinghouse (2017) recommended minimum size of 16 for a sufficiently strong first stage. Corresponding point estimates provided in Table A-11.

## Appendix

**Table A-1. Examination of treatment receipt by compliance**

Panel A. Comparison of no-shows and control group compliers

| | School transformation coaching | | | | Instructional coaching | | | |
|---|---|---|---|---|---|---|---|---|
| | No coaching | | Got coaching | | No coaching | | Got coaching | |
| | Sum | Percent | Sum | Percent | Sum | Percent | Sum | Percent |
| No-shows | 3 | 60.0 | 2 | 40.0 | 2 | 40.0 | 3 | 60.0 |
| Control group compliers | 33 | 63.5 | 19 | 36.5 | 24 | 51.1 | 23 | 48.9 |
| Chi-sq | 0.023 | | | | 0.221 | | | |
| p-value | 0.878 | | | | 0.638 | | | |

Panel B. Comparison of always-takers and treatment group compliers

| | School transformation coaching | | | | Instructional coaching | | | |
|---|---|---|---|---|---|---|---|---|
| | No coaching | | Got coaching | | No coaching | | Got coaching | |
| | Sum | Percent | Sum | Percent | Sum | Percent | Sum | Percent |
| Always-takers | 0 | 0.0 | 6 | 100.0 | 0 | 0.0 | 6 | 100.0 |
| Treatment group compliers | 4 | 6.9 | 54 | 93.1 | 5 | 8.9 | 51 | 91.1 |
| Chi-sq | 0.441 | | | | 0.583 | | | |
| p-value | 0.506 | | | | 0.445 | | | |

Calculations from principal survey data. Survey question for column with school transformation coaching was, “Since January 2016, did you meet in-person, one-on-one with a school transformation coach or someone who has provided you with deliberate, sustained assistance designed to help you learn or figure out how to improve your current school?” Response options were (1) Yes, I received School Transformation Coaching from NC DPI, (2) Yes, I received advice/guidance/coaching from a source other than NC DPI, and (3) No. Responses of (1) and (2) were both coded as having received coaching, while a response of (3) was coded as not having received coaching. The response rate for this question was 81% for principals of schools assigned to treatment and 70% for principals of schools not assigned to treatment. Survey question for column with instructional coaching was “Have any of your teachers received in-person instructional coaching within your school building since January 2016?” Response options were (1) Yes, (2) No, and (3) I don’t know. A response of (1) was coded as having received coaching, a response of (2) was coded as not having received coaching, and a response of (3) was coded as missing. The response rate for this question 78% for principals of schools assigned to treatment and 64% for principals of schools not assigned to treatment.

**Table A-2. Tests for validity of implementation groupings**

Panel A: Comprehensive Needs Assessments and CNA unpackings

| | (1) CNA timing | (2) Unpacking timing | (3) Unpacking presence |
|---|---|---|---|
| 2014 or 2015 | -3.667 | | |
| | (2.137) | | |
| Spring 2016 | -0.447 | | |
| | (1.976) | | |
| 2016-17 school year | 2.221 | | |
| | (2.085) | | |
| 2014 or 2015 | | -4.360* | |
| | | (1.922) | |
| Summer 2016 | | -0.541 | |
| | | (1.708) | |
| 2016-2017 during SY | | -0.0667 | |
| | | (2.306) | |
| Unpacking occurred | | | -1.307 |
| | | | (1.496) |
| Constant | 27.38*** | 28.00*** | 27.82*** |
| | (1.582) | (1.122) | (1.209) |
| $R^2$ | 0.112 | 0.0752 | 0.0104 |
| Obs | 75 | 75 | 75 |

Estimates from regressions of 2015 performance composite on CNA timing group, unpacking presence group, and unpacking timing group, respectively. Reference categories are no CNA/pre-2014 CNA group, no unpacking group, and no unpacking/pre-2014 unpacking group, respectively. Standard errors in parentheses.

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

Panel B: Coaching dosage

| | (1) Total coaching | (2) Instructional coaching | (3) School transformation coaching |
|---|---|---|---|
| Bottom quartile | 3.416 | -2.017 | 0.675 |
| | (1.743) | (1.729) | (1.766) |
| Highest quartile | -1.509 | -4.044* | -2.373 |
| | (1.650) | (1.644) | (1.680) |
| Constant | 26.59*** | 28.58*** | 27.46*** |
| | (0.969) | (0.998) | (1.020) |
| $R^2$ | 0.0834 | 0.0788 | 0.0384 |
| Obs | 75 | 75 | 75 |

Estimates from regressions of 2015 performance composite on coaching dosage group. Reference category is middle 50% of schools. Standard errors in parentheses.

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

**Table A-3. TOT estimates within alternative bandwidths and full sample *(outcome=test score)***

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 *F*-stat | 160.28 | 156.25 | 76.91 | 3.69 | 31.58 | 30.69 |
| 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 *F*-stat | 158.76 | 144.24 | 48.86 | 5264.95 | 29.59 | 62.88 |
| 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 *F* statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 16 for a sufficiently strong first stage.

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

**Table A-4. TOT estimates by subject *(outcome=test score)***

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 *F*-stat | 79.92 | 34.57 | 199.37 | 79.57 | 29.48 | 212.87 |
| 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 *F*-stat | 79.92 | 22.94 | 153.76 | 54.91 | 19.10 | 154.75 |
| 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 *F*-stat | 901.80 | 51810.86 | 924.16 | 1109.56 | 6.9224e+29 | 1455.42 |
| 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 F statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 16 for a sufficiently strong first stage.

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

**Table A-5. 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 *F*-stat | 72.76 | 20.16 | 122.99 | 47.89 | 16.00 | 132.48 |
| 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-6. TOT estimates by school level *(outcome=test score)***

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 *F*-stat | 3.13 | 0.50 | 6.76 | 2.62 | 0.38 | 7.18 |
| 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 *F*-stat | 303.46 | 2.43049e+35 | 381.81 | 442.26 | 5.184e+35 | 598.29 |
| 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 F statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 16 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-7. 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 *F*-stat | 78.50 | 76.91 | 12.04 | 11.02 | 6.10 | 3.84 |
| 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 *F*-stat | 80.46 | 78.68 | 18.06 | 390.46 | 6.25 | 4.41 |
| 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 *F*-statistics on the treatment indicator smaller than the What Works Clearinghouse (2017) recommended minimum size of 16 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-8. 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-9. Placebo estimates from fuzzy RD within optimal CCT bandwidth, 2016 *(outcome=test score)***

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-10. School demographics by treatment year**

| | 2016 | | | 2017 | | |
|---|---|---|---|---|---|---|
| | Treat | Control | p-value | Treat | Control | p-value |
| ED percent | 65.34 | 68.62 | 0.561 | 68.32 | 65.37 | 0.780 |
| Minority percent | 77.60 | 74.34 | 0.702 | 78.20 | 67.41 | 0.340 |
| Black percent | 48.72 | 48.89 | 0.988 | 49.45 | 42.41 | 0.650 |
| Hispanic percent | 16.72 | 20.11 | 0.583 | 17.73 | 20.33 | 0.710 |
| ADM | 418.23 | 433.38 | 0.885 | 399.75 | 428.53 | 0.808 |

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

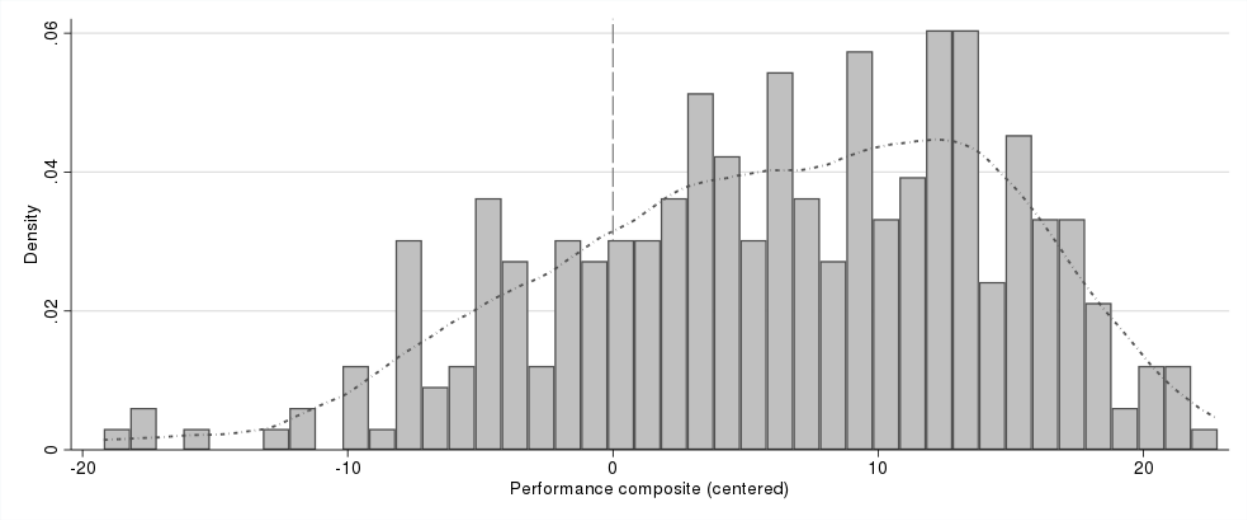
**Table A-11. Fuzzy RD results by CNA timing *(outcome=test score)***

| | 2016 | | 2017 | |
|---|---|---|---|---|
| | (1) Full sample | (2) CCT | (3) Full sample | (4) 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 F-statistics are greater than What Works Clearinghouse (2017) recommended minimum size of 16 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.

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

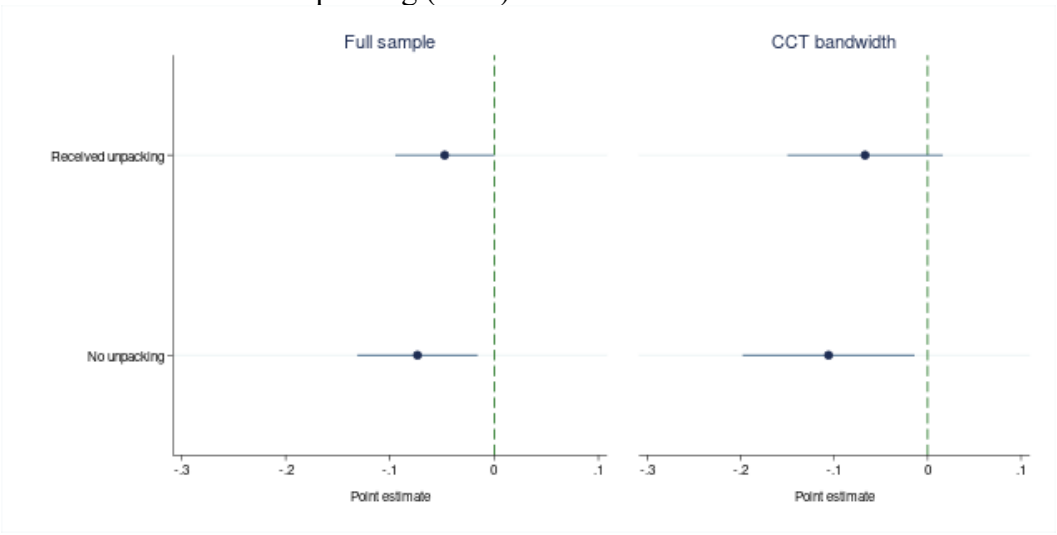
**Figure A-1. Graphical integrity of the forcing variable**



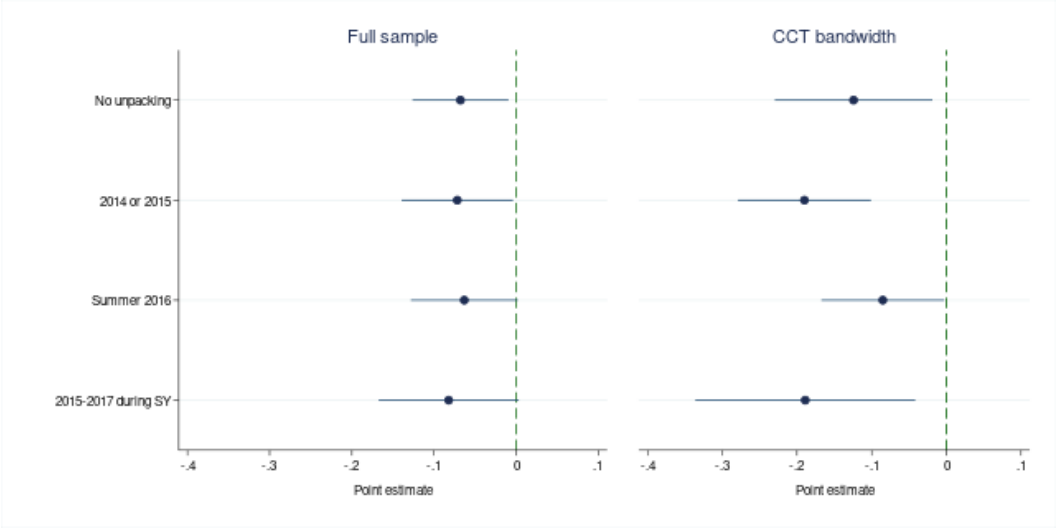

NOTE: Bin width is 1. Includes all eligible schools.

**Figure A-2. Estimated effects by presence of CNA unpacking**

Panel A: Presence of Unpacking (2017)



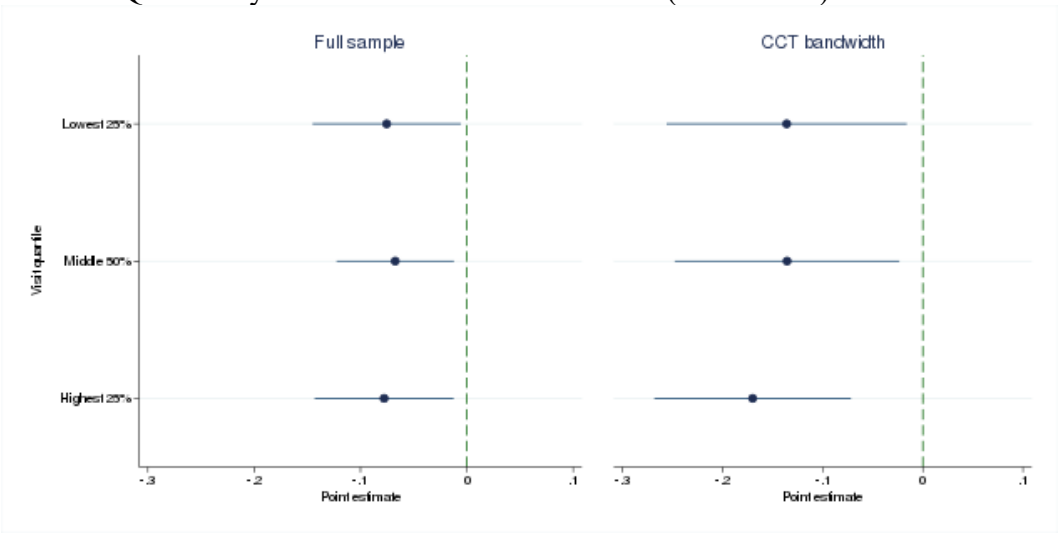

Panel B: Timing of Unpacking (2017)



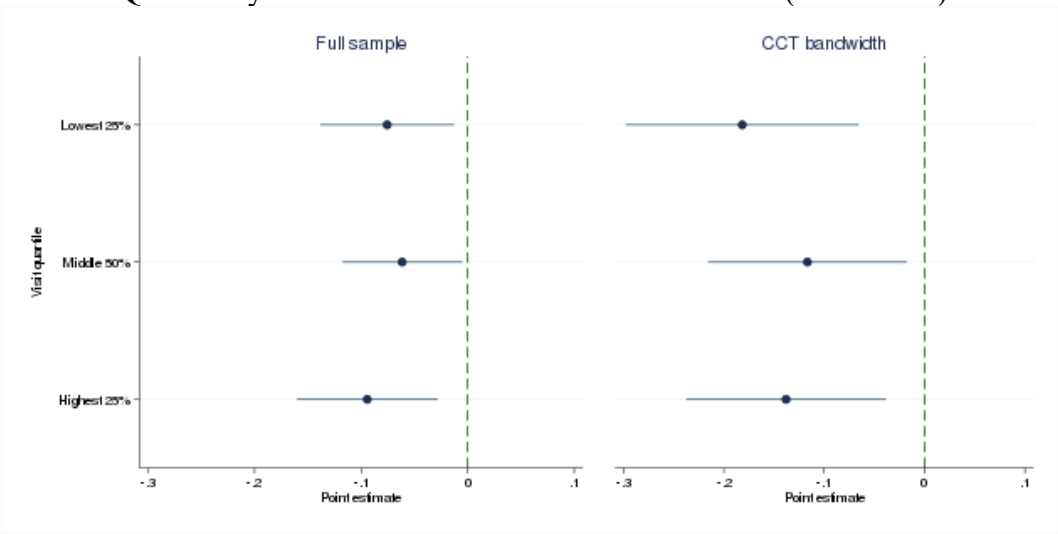

NOTE: 2SLS estimates from fuzzy RD using triangular kernel with separate treatments by CNA unpacking presence (panel A) and timing (panel B). All models include lagged score and subject on the right side, with math as the reference category. All first-stage *F*-statistics are greater than What Works Clearinghouse (2017) recommended minimum size of 16 for a sufficiently strong first stage. Preferred CCT bandwidths from fuzzy test score models. Standard errors clustered at the school level.

**Figure A-3. 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 *F*-statistics are greater than What Works Clearinghouse (2017) recommended minimum size of 16 for a sufficiently strong first stage. Preferred CCT bandwidths from fuzzy test score models. Standard errors clustered at the school level.